João Filipe Ferreira Curado e Silva

Transcription, splicing and chromatin structure

Tese de Candidatura ao grau de Doutor em Biologia Básica e Aplicada submetida ao Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto.

Orientador: Roderic Guigó
Categoria: Professor catedrático
Afiliação: Centre for Genomic Regulation
Universitat Pompeu Fabra
Barcelona

Co-orientadora: Alexandra Moreira
Categoria: Investigadora Principal
Afiliação: Instituto de Biologia Molecular e Celular
Universidade do Porto
Porto

ABSTRACT

In the last few years, chromatin states have been characterized with great detail and associations between histone modifications, transcription and splicing have been reported. Here we took advantage of the huge quantity of data being generated by big international consortiums to undertake different approaches and study these connections in a genome-wide level. First we explored transcription of developmental regulated genes in the context of chromatin. We found that these genes are expressed in the absence of canonically active histone modifications and possibly under stronger regulation of transcription factors. In this section we also report that strong chromatin marking is associated with stable expression and splicing. Since splicing, the process of intron removing from pre-mRNAs, can happen co-transcriptionally, in the second part of this thesis we aimed to quantify how often this phenomenon was happening. Our results demonstrate that the majority of the splicing events occur while RNA polymerase II is still transcribing. This evidence led us to investigate, in a genome-wide level, the connection between splicing and chromatin. We found that, despite having very few exons being differentially included between a panel of human cell lines we could identify a subset of them whose inclusion level is strongly correlated with the presence of some chromatin features. In these exons, increased levels of H3K9ac, H3K27ac, H3K4me3, DNase I hypersensitive and transcription factor binding are associated with higher exon usage. We also found that when in conditions of higher inclusion these exons are also in close proximity (either in linear of in the 3D space) of active promoters. These observations suggest a functional role for histone modifications and other chromatin features that activate transcription in the regulation of splicing of exons in their physically proximity.

RESUMO

Nos últimos anos foram caracterizados em grande detalhe vários estados de cromatina e foram descobertas associações entre modificações de histonas, transcrição e splicing. Neste projeto aproveitamos a enorme quantidade de dados gerados por grande consórcios internacionais para levar a cabo várias estratégias e estudar estas relações à escala do genoma completo. Na primeira parte explorámos a transcrição de genes regulados durante o desenvolvimento sob o contexto da cromatina. Descobrimos que estes genes são expressos na ausência de modificações histónicas canonicamente associadas com expressão ativa, estando possivelmente sob uma mais forte regulação por parte de factores de transcrição. Nesta secção também reportamos que uma forte modificação da cromatina se encontra associada com estabilidade de expressão e de splicing. Uma vez que o splicing, o processo da remoção de intrões do pre-mRNA, pode acontecer co-transcricionalmente, na segunda seção desta tese decidimos quantificar a prevalência deste fenómeno. Os nossos resultados demonstram que a maioria dos eventos de splicing ocorrem durante o período em que a RNA polimerase II está a transcrever. Esta evidência levou-nos a investigar, à escala do genoma completo, a conexão entre splicing e cromatina. Descobrimos que, apesar de poucos exões terem os níveis de inclusão regulados entre várias linhas celulares humanas, era possível identificar um subgrupo de exões com o nível de inclusão fortemente correlacionado com a presença de alguns características da cromatina. Neste subgrupo de exões, altos nível de H3K9ac, H3K27ac H3K4me2, sensibilidade a DNase I e presença de factores de transcrição estão associado a um aumento do nível de inclusão. Também descobrimos que, aquando em condições de maior nível de inclusão este subgrupo de exões está em proximidade (tanto a nível linear como no espaço 3D) de promotores ativos. Estas evidências sugerem um papel funcional das modificações de histonas e outras características da cromatina que ativam a transcrição, na regulação do processo de splicing de exões na sua proximidade.

ACKNOWLEDGMENTS

I want to thank everybody that has been with me, encouraged me, inspired me and in any way helped me during this incredible and really important journey that has been my PhD. Without all these friends I wouldn't have made it for sure.

The first one I owe a big thank you is Roderic, my thesis supervisor, who is a true inspiration not only as a scientist but also as a person and as a boss. It has been a great pleasure and the biggest educational experience in my life to be under his supervision. I thank him for giving me the opportunity to work in this very interesting and challenging (also frustrating…) project. With him I learned what being a scientist means.

Past and current members of Guigó's group were wonderful and a really pleasure to work with. Everything is easier when you do it with friends. Thank you guys! I need to give a special thanks to Hagen who accepted me almost has his trainee when I first arrived in Barcelona and to who I can still trace back almost everything in my (good) working practices. I also thank Alessandra, who was so kind to share her knowledge (and code) with me all the time. Julien and Rory, thanks for being great buddies. Thanks Barbara for being such a lovely assistant. I also need to thank Romina for running everything in the background. Sarah, Dmitri, Emilio and all the others, thanks for everything. And of course, thanks Francisco and Pedro, the Portuguese connection in the lab that made it so much easier to feel at home.

I also need to thank Camilla for the collaboration in the chromatin/splicing (long…) project, and Juan for the supervision! Silvia, Enrique and Montse, thank you for all the meetings and scientific discussions (It was also a very long collaboration…). And a very "technical thanks" to Monica who helped me with the figures in this thesis.

I also want to give a big thank you to all the friends that I met during this period of my life. Michael and David who really made me see life in a different and better way, Lucia that was always by my side and

always ready to help. The Ducs, The FC Lusitanos, the Gentlemen, the NPC team, the Incubakers and all the people that made me have such an amazing time in Barcelona, thank you guys!!

From Portugal I have to thank everybody from my PhD program, GABBA, in particular the 12th edition crew with whom I shared some of the best moments of my life, Catarina Carona for all the help with bureaucracy etc., Professora Alexandra for the co-supervision and all the GABBA mentors for giving me this amazing opportunity. Before I finish, of course I want to give a very special shout to my family and old friends that were always so supportive and gave me so much energy to continue.

A big thanks to all the above and all the others that I didn't mention and that somehow have been there in these last few years.

Obrigado!

TABLE OF CONTENTS

CHAPTER 1- INTRODUCTION

A. The basics of Pol II transcription

Transcription is the first and most highly regulated step in eukaryotic gene expression (Nechaev and Adelman). Polymerase II (Pol II) transcription is initiated by activator proteins binding upstream the transcription start site (TSS), at the core promoter. A series of protein-protein interactions result in the recruitment of Pol II and various transcription factors (TF). DNA at the TSS starts to be unwounded and the active site of Pol II is positioned in the single-stranded DNA template. Early elongation starts after the phosphorylation of serine 5 (Ser5) of the carboxyl-terminal domain (CTD) of RPB1, the largest Pol II subunit. This C-terminal domain contains 52 repeats of the amino acid sequence Tyr-Ser-Pro-Thr-Ser-Pro-Ser, offering multiple possible phosphorylation sites. The serine residues are referred to as Ser2, Ser5, Ser7 due to their position in this sequence (Corden et al.). Ser2 position is further phosphorylated, helping in the recruitment of important factors for transcription elongation and giving start to the efficient elongation step. During the transcription cycle, Pol II CTD serves as a recruitment platform for a large number of factors required for productive elongation and messenger RNA (mRNA) processing, including histone modifiers and splicing factors (Smolle and Workman).

B. Transcription and chromatin

B.1 Nucleosomes

The nucleosome core particle consists of approximately 147 base pairs (bp) of DNA wrapped around a histone octamer – 2 copies of each of the core histones H2A, H2B, H3 and H4 – forming one of the most stable protein-DNA complexes known. Nucleosome positioning and depletion is determined by DNA sequences, histone variants incorporation and transcription factors. Although their main function is packaging, nucleosomes possess many dynamic proprieties, tightly regulated by various protein complexes. Depending on the context, nucleosomes can inhibit or facilitate transcription factor binding. Most transcribed genes have reduced nucleosome occupancy over their promoters when compared to the rest of the genome (Müller and Tora).

B.2 Histone modifications and transcription

Core histones are globular proteins with a long unstructured N-terminal tail that can undergo a variety of post-translational modification in multiple residues (Figure 1). Such modifications influence chromatin structure, the recruitment of other proteins to chromatin (Kouzarides) and transcription (Li, Carey, and Workman).

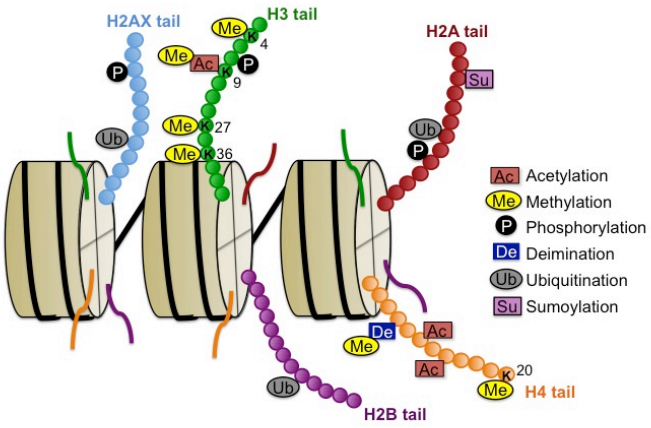


Figure 1. Histones can undergo a variety of post-translational modifications such as acetylation, methylation or phosphorylation in some specific residues of their long N-terminal tails.

In the last few years, histone modification states associated with active transcription or silenced chromatin have been characterized with great detail (Cheng et al.; Filion et al.). Acetylation (ac) of histone 3 and histone 4 (H3 and H4) and methylation (me) of lysine (K) 4 and 36 of H3 are associated with active transcription and often referred as euchromatin modifications. In the other hand, trimethylation of Lysine 9 or 27 are normally found in inactive genes or repressed regions and can be termed heterochromatin modifications (Li, Carey, and Workman). Distribution patterns of histone modifications and their correlation with transcription rates can be seen in Figure 2.

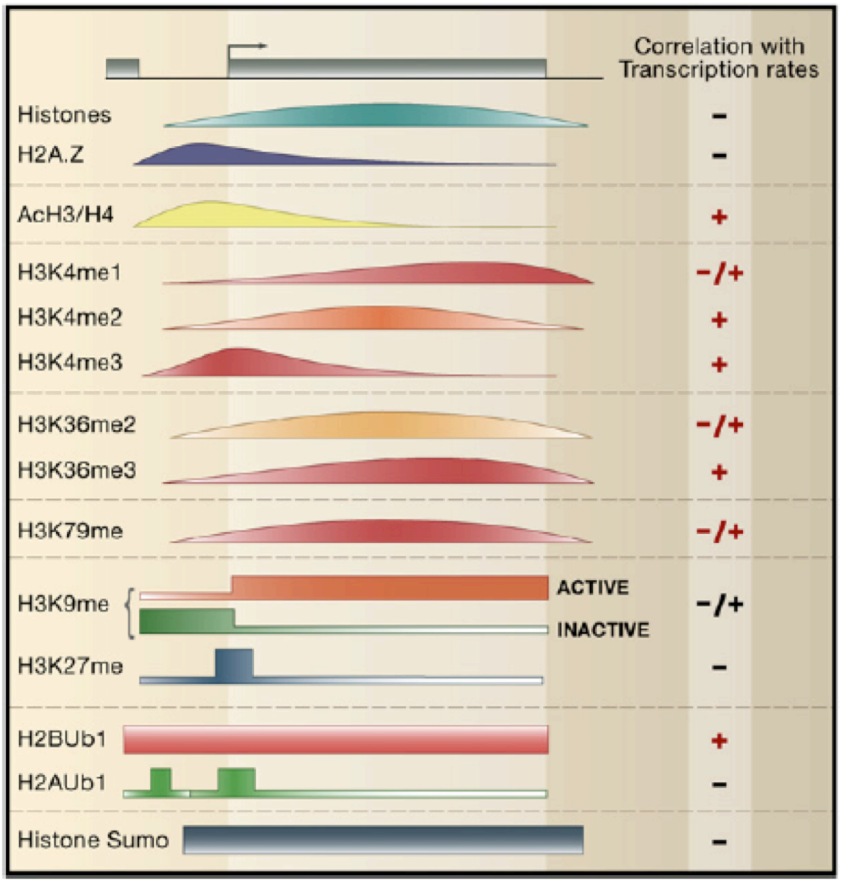


Figure 2. Distribution of histone modifications on an arbitrary gene and their correlation signs with transcription rates. The curves represent the patterns that are determined via genome-wide approaches. Adapted from Li et al 2007.

Nucleosomes are very strong barriers for Pol II transcription (Kireeva et al.). Although striking in detail, these well-defined landscapes and associations may also be a consequence of the struggle to maintain chromatin structure in the transcribed regions during Pol II traversing. When transcribing into nucleosome templates, Pol II pauses at certain sites with stronger histone-DNA contact. To overcome this barrier, ATP-dependent chromatin-remodeling complexes are known to act, allowing transcription to go through (Carey, Li, and Workman).

B.3 Transcription in the absence of H3K4me3

Modifications occurring in the DNA and histone proteins are seen to correlate and maybe even control gene expression by establishing and maintaining specific chromatin states (Delest, Sexton, and Cavalli). Nevertheless, recent genome-wide studies reported the lack of H3K4me3 at the TSS of a fraction of the expressed genes in *Drosophila Melanogaster* (Nègre et al.). Even more striking was the finding that cells from the wing imaginal disc of *Drosophila*, carrying a non-methylable residue instead of lysine 4, could activate gene expression and respond to the developmental signaling pathway (Hödl and Basler). Although widespread it seems that the presence of H3 protein species is not essential in fly and transcriptional regulation can occur in the complete absence of H3K4me3.

C. Pre-mRNA splicing

C.1 Splicing and alternative splicing

The existence cycle of RNA molecules begins with transcription and ultimately ends in degradation. Messenger RNA, the RNA molecule that carries the genetic information from DNA to the ribosome, may also be processed, depending on the species. After transcription, precursor RNA (pre-mRNA), undergoes different modifications such as 5' cap addition, splicing, editing and polyadenylation. Splicing is the process by which introns are removed from the nascent pre-mRNA and exons are joined, generating the functional mRNA. The majority of pre-mRNAs contain exons that can be alternatively included or excluded into the mature mRNA in a process called alternative splicing. Exons can be extended or skipped and introns can be retained (Figure 3). Alternative splicing is a highly regulated mechanism that endows a single gene with the capacity to generate multiple RNA molecules, either encoding proteins with different functions (Kelemen et al.) or different intergenic long non-coding RNAs (lncRNAs) (Derrien et al.). Being splicing a major contributor to protein diversity in metazoans, it is estimated that 95% of multi-exon genes undergo alternative splicing, sometimes in a tissue-specific manner and/or under specific cellular conditions (Pan et al.; Wang and Burge).

Sequence elements at or near intron/exon borders are recognized by the spliceosome, a large and complex molecular machine in charge of intron removal. This complex is composed of five small nuclear ribonucleoprotein particles (snRNPs) and more than 150 auxiliary proteins (Wahl, Will, and Lührmann). The binding of additional trans-acting factors, members of the SR protein family (Serine and Arginine rich) and hnRNPs, to cis-regulatory sequence elements (splicing silencers and enhancers) located in intronic or exonic regions of the pre-mRNA facilitates or prevents spliceosome assembly on regulated splice sites (Chen and Manley). In this classical view, the combination of all these features, acting

in combination on the transcribed pre-mRNA determines the splicing outcome (Barash et al.).

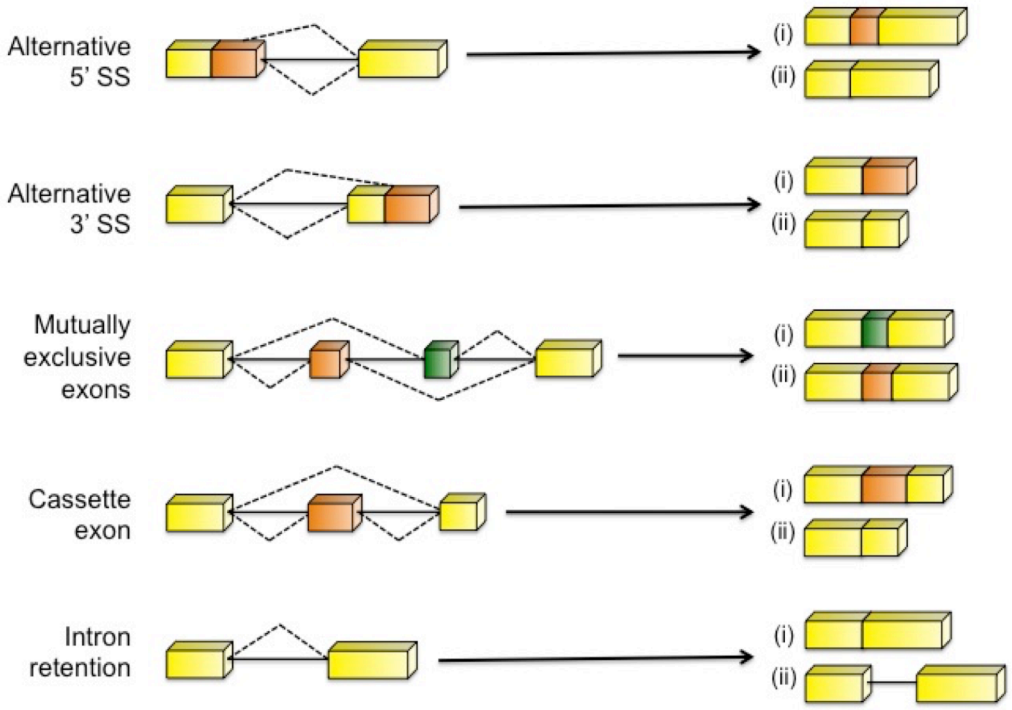


Figure 3. Examples of alternative splicing events. Exons are represented by blocks and introns by lines. In orange and green there are the exons or portions of exons that can be alternatively included or skipped. In the right side there are the 2 possible transcripts coming from each example gene, on the left.

C.2 Importance of alternative splicing

The overall function of alternative splicing is to increase the diversity of the mRNA expressed from the genome. Genome-wide studies indicate that coding alternative exons predominantly encode coiled regions on the outside of the protein. Alternative exons generally influence local regions on the protein surface without changing or disrupting the general structure of the protein (Wang et al.; Romero et al.). Evolutionary analysis

support this notion, as human-mouse comparison shows that alternative exons preserving the open reading frame (ORF) are conserved between these species (Zhang, Krainer, and Zhang).

This doesn't mean that alternative splicing doesn't have a function. In matter in fact, the amount of alternative splicing increases from invertebrates to vertebrates (Kim, Magen, and Ast) and in the latter it's also more rich in genes specific of the immune and nervous system (Modrek and Lee). This suggests that alternative splicing is more important in more complex systems, where information must be processed differently at different times or a higher degree of variety is needed.

In the other hand, given the generality of alternative splicing events it is not surprising that aberrant regulation of the process can lead to disease. It is estimated that regulation of splicing could be involved in 15% of genetic diseases (Krawczak, Reiss, and Cooper). A single polymorphism in an exonic enhancer can cause mis-splicing of exon 18 of BRCA1 and lead to cancer (Liu et al.). Other associations with diseases are known, either with disrupting of the splicing of specific genes (Bechtel et al.) or with general deregulation, in cancer cells (Klinck et al.; Venables et al.). Localization and enzymatic proprieties of the proteins can also change due to changes in alternative splicing outcome (Kelemen et al.).

C.3 Co-transcriptional splicing

Splicing, from the beginning, was defined as a post-transcriptional process, well documented *in vitro* and *in vivo* (Green, Maniatis, and Melton). With the realization that introns could be removed while nascent transcripts were still tethered to the DNA a new dimension of alternative splicing regulation arose (Beyer and Osheim). Recent evidence, in which this research project took part in, changed the paradigm and it is now commonly accepted that, in most of the cases, transcription and pre-mRNA splicing are an integrated process (Ameur et al.; Dye, Gromak, and

Proudfoot; Khodor, Menet, et al.; Khodor, Rodriguez, et al.; Listerman, Sapra, and Neugebauer; Pandya-Jones and Black; Tilgner, Knowles, et al.; Vargas et al.).

We know now that splicing is mostly co-transcriptional but we're still trying to understand how both processes are coupled and how can splicing and transcription machinery interact. Three aspects, somehow related, of co-transcriptional processing can influence alternative splicing regulation. First, a kinetic competition between splice sites where a weaker splice site can be recognized and used under pausing or slow elongation rates as opposed to the stronger splice site, always used, independently of the rate of elongation (Roberts et al.). In this model, inclusion levels anti-correlate with polymerase speed (Sebastián Kadener et al.; Mata et al.).

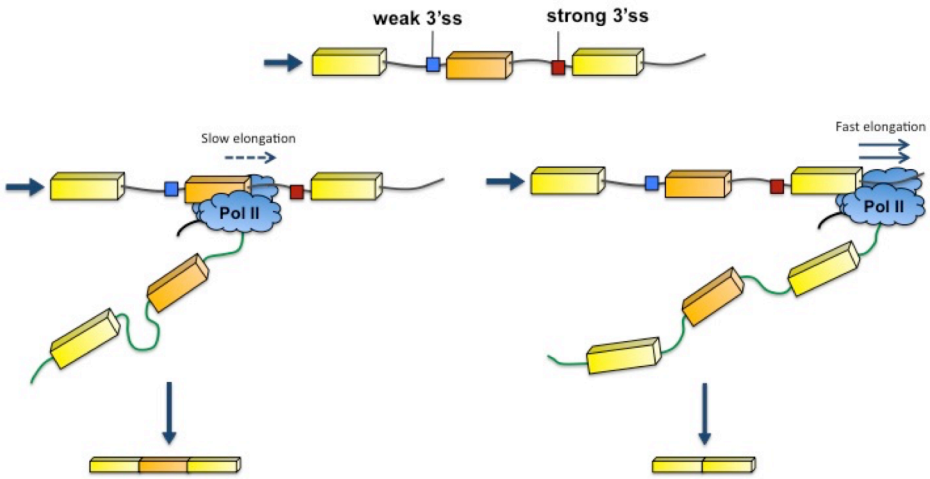


Figure 4. Model of alternative splice site events regulated by RNA polymerase II elongation rates. In the model, an alternative exon (in dark yellow) preceded by a weak splice site is more included when elongation rate of RNA polymerase is low (left) and is more skipped when elongation rate is faster (right). Slower elongation rates provide more time for excision of the intron upstream of the alternative exon before its splice site enters in competition with the splice site of the downstream exon.

Second, if splicing can take place while RNA and DNA molecules are still attached to Pol II, the latter can play a role in splicing regulation. In particular the CTD of Pol II is known to interact with a variety of transcription and splicing regulators. For example, transcription termination and 3' end processing depend on the CTD-domain of Pol II (McCracken et al.), hinting that these processes not only occur in physical proximity to transcription, but that there could be a functional connection. Third, given the intimate relationship between transcription regulation and chromatin structure, already for some time, models for splicing regulation by chromatin had been proposed. Recently, with the advent of massively genome wide sequencing, the resolution for nucleosome and histone modification profiling has substantially increased allowing us for the first time to compare signal in exons versus introns, for example. Cases of direct chromatin–splicing interactions have been reported: a subunit of the chromatin remodeling complex SWI/SNF regulates splicing (Batsché, Yaniv, and Muchardt) and many different histone modifications were already implicated in splicing regulation (Alló et al.; Nogues et al.; Schor et al.; Sims et al.; Luco, Pan, et al.).

C.4 Nucleosome organization on exons

Vertebrates' exon size, in contrast with the much longer and variable intron size, seems to be under strong evolutionary constraint remaining at a small range around 140 base pairs (Wang and Burge). Already in 1991, Beckmann and Trifonov noticed a periodic pattern between splice sites, compatible with nucleosome phasing (Beckmann and Trifonov). They hypothesized the presence of a nucleosome could protect splice sites from mutational hazards.

Micrococcal nuclease digestion (MNase) of chromatin followed by high-throughput DNA sequencing revealed that, indeed, regions protected

from nuclease digestion are enriched in exons over introns. This confirmed that nucleosomes are preferentially positioned in exons and supports the protective role of nucleosomes, mentioned above. This is a phenomenon evolutionary conserved (Hodges et al.; Nahkuri, Taft, and Mattick; Schwartz, Meshorer, and Ast; Spies et al.; Tilgner, Nikolaou, et al.). Several other correlative features argue in favor of a functional role of nucleosomes in splice site recognition and exon definition. Isolated exons flanked by long introns have higher nucleosome occupancy than clustered exons with small introns (Spies et al.). Exons with weaker splice sites have stronger nucleosome occupancy, suggesting that well-positioned nucleosomes can be key in the recognition of weak splicing signals (Tilgner, Nikolaou, et al.; Schwartz, Meshorer, and Ast). Nucleosome occupancy in exons is independent of the transcription level (Tilgner, Nikolaou, et al.) and is enhanced in highly included versus highly excluded exons (Spies et al.). Pseudoexons (intronic sequences with well-defined splice site sequences that are not included in mature transcripts) show, inversely, a depletion of nucleosomes (Tilgner, Nikolaou, et al.). And lastly, nucleosome enrichment is present, independently of sequence conservation or GC- content although the strength of nucleosome occupancy is correlated with GC-content levels (Nahkuri, Taft, and Mattick; Schwartz, Meshorer, and Ast; Tilgner, Nikolaou, et al.).

C.5 Histone modifications on exonic regions

Genome-wide experiments, in particular chromatin immunoprecipitation followed by sequencing (ChIPSeq), allowed the study of histone modifications and variations in genomic regions. Preferential accumulation of particular histone modifications was found in exons over introns (Andersson et al.; Kolasinska-Zwierz et al.; Schwartz, Meshorer, and Ast; Spies et al.). H3K36me3 (and H3K79me1 in a less pronounced way) levels are known to increase towards the 3’ end of genes and to

correlate with expression (Figure 2). However, inside the genes, the profiling revealed enrichment in exons versus introns, with stronger signal in constitutive ones. This was initially discovered in *C. Elegans* and later extended to human and mouse) (Andersson et al.; Dhami et al.; Kolasinska-Zwierz et al.). Internal exons are also marked by H3K27me2 while H3K27me1 and H3K4me1 signal appears enriched only in internal exons flanked by long introns (Spies et al.). Introns were also described as carrying monoubiquitination of histone 2B marking (Dhami et al.; Shieh et al.).

When looking at exonic regions, one needs to be careful by the increased nucleosome occupancy. Although in some cases is difficult to distinguish histone modification enrichment from nucleosome enrichment, some studies showed that the enrichment of histone modifications present in exons could not be explained by the consequence of having a higher density if nucleosomes. H3K36me3, H3K4me3 and H3K27me2 maintain their enriched profile in exons over introns even after nucleosome normalization (Dhami et al.; Spies et al.; Kolasinska-Zwierz et al.).

Since histone modifications profoundly influence chromatin accessibility (Kouzarides), they have been hypothesized to influence splice site recognition and usage too.

D. Intragenic chromatin organization and splicing

As splicing knowledge grew, it became clear that regulation and splice site selection is a far more complex process than it was anticipated. RNA-binding factors and RNA Pol II elongation rate are not sufficient to fully explain the faithful regulation of alternative splicing. RNA-binding motifs are not conserved between genes and even when they are transcribed with errors, they can still often recruit the appropriate set of splicing factors (Fox-Walsh and Hertel). Similarly, although Pol II elongation rate are known to affect splicing outcome in different scenarios (Mata et al.; Muñoz et al.) it is still unclear if this is a commonly used mechanism *in vivo*. Taken the observations in Section C all together, it was natural that many authors started raising the possibility of having chromatin organization as a key regulator in RNA processing and more specifically, in alternative splicing.

The first evidence of an interplay between chromatin structure and splicing came from the finding that fibronectin exon 33 inclusion or exclusion was sensitive to replication-mediated chromatinization status of the plasmid and to histone deacetylase inhibitor TSA (Sebastian Kadener et al.; Nogues et al.). The recent advent of methods to study chromatin structure in a genome-wide scale revealed association between histone modifications and transcript diversity (Podlaha et al.) and many specific connections that essentially can be categorized in two, not mutually exclusive, models (Figure 5): The “kinetic model” that proposes regulation of splicing outcome by using nucleosome position and/or histone modifications as a proxy for modulating RNA polymerase II elongation rate (Kornblihtt et al.) and the “recruitment model” that argues for direct recruitment of splicing factors on the nascent pre-mRNA by histones and their modifications in a chromatin-splicing adaptor system (Luco, Pan, et al.).

All together it seems natural to speculate that chromatin regulators, able to read the “histone code”, may be helping splicing machinery to

locate and access pre-mRNA and therefore, regulating splicing in addition to all the long time known players.

D.1 Chromatin, elongation rate and splicing

One of the mechanisms by which the transcription machinery can influence splicing is the called “kinetic coupling model”. In this model, the rate of Pol II elongation influences the splicing outcome by affecting the pace at which splice sites and regulatory sequences emerge and are made available in the nascent pre-mRNA, during transcription. This model became popular when the nature of the promoter, and consequent elongation rate, was discovered to influence alternative splicing outcome (Cramer et al.; Kornblihtt). The result was not a consequence of promoter strength but of some qualitative properties conferred by promoters to the transcription machinery. Several other lines of evidence support the idea that RNA Pol II elongation can affect alternative splicing through kinetic coupling. Inhibiting histone deacetylation and therefore promoting “open chromatin” by trichostatin A (TSA) leads to more skipping of the exon 33 (E33) of *fibronectin* (Sebastian Kadener et al.). On the other hand, drugs inhibiting elongation, like DRB (Sebastian Kadener et al.; Nogues et al.), flavopiridol, or camptothecin (de la Mata, Lafaille, and Kornblihtt) favor inclusion. However, the strongest evidence for a kinetic role of Pol II elongation comes from a slow mutant Pol II, which increases the same exon 33 usage in human cells when compared to the normal rate polymerase (Mata et al.).

Another very interesting finding came from the study of the DNA-binding protein CCCTC-binding factor (CTCF), a genomic insulator. CTCF was described to create a “roadblock” to Pol II elongation downstream the exon 5 of *CD45*, promoting its inclusion (Shukla et al.). Additionally, DBIRD, a two-subunit complex that interacts with RNA polymerase II and nascent ribonucleoprotein complexes (mRNPs), which promotes RNA Pol II elongation through A-T rich regions (particularly difficult for Pol II to

transcribe) was reported to facilitate skipping of alternative exons within these regions (Close et al.).

Since nucleosomes represent physical barriers for the progression of RNA polymerase complex they also have the capacity to modulate transcription elongation rates. In recent studies, Pol II was reported to pause at nucleosomes (Churchman and Weissman; Kwak and Lis) and at the 3' end of introns (Kwak and Lis) which is consistent with the higher nucleosome density in exons, described before. Moreover, *Brahma* (Brm) and Brg1, protein components of the SWI/SNF chromatin remodeling complex that can influence nucleosome occupancy, were found to associate with regulators of the spliceosome, slowing down polymerase II and favoring inclusion of alternative exons (Batsché, Yaniv, and Muchardt). This chromatin binding is in part mediated by a bromodomain that recognizes acetylated histones (Agalioti, Chen, and Thanos). Histone acetylation was also implicated in splicing regulation in a more physiological context by Schor et al., in 2009. Upon neuron membrane depolarization, acetylation of H3K9 is increased around exon 18 of the Neural Cell Adhesion Molecule (NCAM) gene but not in its promoter. Exon 18 is also reported as more skipped while Pol II elongation rate is increased (Schor et al.). All this evidence is compatible with the kinetic model of splicing regulation (Figure 5a).

On the other hand, a recent study showed that a slow Pol II could also induce skipping of exons. In CFTR gene, skipping of exon 9 is promoted by a slow elongation rate through the recruitment of the negative splicing factor ETR-3 onto de uridine and guanosine (UG) repeat located at the neighbor intron (Dujardin et al.).

D.2 Chromatin-splicing adaptor system

In the chromatin context, adaptor molecules are molecules that recognize specific epigenetic marks and interact with factors and enzymes

involved in chromatin modifications. In transcription regulation the role of adaptor molecules is very well documented (Li, Carey, and Workman; Ries and Meisterernst). The protein MRG15 was one of the first published evidences of a molecule with a similar role in splicing regulation. MRG15 recognizes H3K36me3 and interacts with the splicing regulator Polypyrimidine Tract Binding protein (PTB) regulating splicing of the gene FGFR2 (Luco, Pan, et al.). Exon IIIb of FGFR2 is used in epithelial cells while exon IIIc is included in mesenchymal cells instead. Interestingly these two cell types display differences in histone modification profiles of this region. High levels of H3K36me3 and H3K4me1 in the first, correlating with usage of exon IIIb and high levels of H3K27me3 and H3K4me3 in the second, correlating with usage of exon IIIc. High levels of H3K36me3 along the alternatively spliced region of the gene attract MRG15, which in turn interacts with PTB, recruiting it to the nascent RNA and promoting inclusion of exon IIIb and exclusion of exon IIIc. In contrast, in cell types where H3K36me3 levels are low, the splicing repressor PTB is poorly recruited to the region and, as a consequence, inclusion of PTB-dependent exons favored (Figure 5b). Increasing H3K36me3 levels in the absence of MRG15 protein have no effect on alternative splicing. Importantly, genome-wide analyses uncovered an inverse correlation between alternatively spliced exons affected by MRG15 and the strength of PTB binding sites in the target pre-mRNA (Luco, Pan, et al.) showing that histone modification regulation is particularly important when the intrinsic binding signal is weak, strengthening its effect.

There is evidence of more chromatin-splicing adaptor systems in mammalian cells. H3K4me3 levels play a role in recruitment of the early spliceosome to human cyclin D1 pre-mRNA via binding of the chromatin-adaptor protein CHD1, increasing splicing efficiency (Sims et al.). CHD1 is a chromodomain protein that recognizes H3K4m3 and interacts with SF3a, a sub complex of the U2 snRNP involved in early 3' splice site recognition. CHD1 is also a component of the histone acetyltransferase SAGA complex with which Gcn5 (which binds to acetylated H3) also interacts, recruiting U2 snRNP components to the exon (Gunderson and Johnson). HP1 protein,

a characteristic heterochromatin component, might be another example. Mass spectrometry identified H3K9 trimethylation, HP1 proteins and splicing factors SRp20 and AS/SF2 as interaction partners (Loomis et al.). Co-immunoprecipitation experiments confirmed that HP1β interacts with splicing factors ASF/SF2 in humans (Loomis et al.) and HP1α with hnRNP proteins in *Drosophila* (Piacentini et al.). It was observed that depletion of the HP1α (CBX) isoform is associated with accumulation of unspliced nascent transcripts and deficient recruitment of splicing factors (Smallwood et al.). These results point to a possible role for HP1 as an adaptor between heterochromatin marks and splicing factors although the functional relevance to splice site selection remains to be determined.

a)

b)

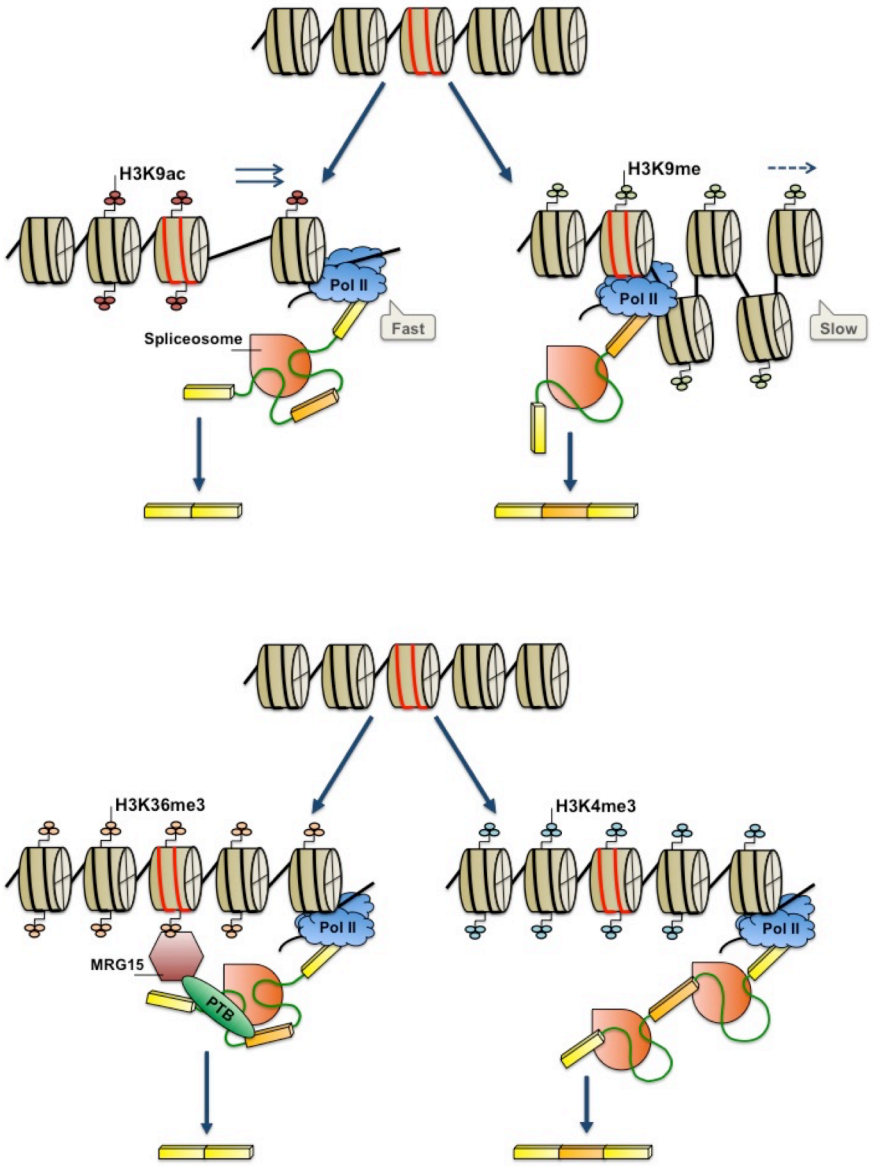


Figure 5. Two known models for splicing regulation by histone modifications. In the **kinetic model (a)** histone modifications (in the example H3K9ac/me) can modulate the speed of polymerase and determine the inclusion/skipping of weak splice sites. In the **recruitment model (b)** an adaptor protein (MRG15 in the example) recognizes H3K36me3 and interacts with a splicing factor (PTB) promoting exon skipping

E. Splicing reaches back

Although the mechanisms postulated so far describe only the effects of chromatin organization on RNA polymerase elongation and recruitment of splicing factors there is also recent evidence of an interaction loop where the recruitment of the splicing machinery can also influence chromatin structure/dynamics and transcription and therefore modulate splicing through these effects (de Almeida and Carmo-Fonseca).

Splicing was first suggested to enhance transcription by Fong and Zhou, who found that spliceosomal snRNPs interact with the transcription elongation factor TAT-SF1. TAT-SF1-snRNP complexes can stimulate transcription elongation *in vitro* in a splice site dependent manner (Fong and Zhou). Splice sites were also reported to cause stalling of polymerase when mutated (Martins et al.) and in terminal exons in yeast (Carrillo Oesterreich, Preibisch, and Neugebauer). In *Saccharomyces cerevisiae*, RNA polymerase II was found to accumulate transiently around 3' end of introns. This apparent pausing coincides with splicing factor recruitment and it terminates upon appearance of spliced products (Alexander et al.). These findings suggest that intron removal is required for transcription termination and pausing, in turn, could be part of a proofreading mechanism for correct mRNA production.

Splicing regulatory Hu proteins can also influence transcription changes by modulating chromatin structure (Zhou et al.). Hu proteins are recruited to pre-mRNA and can induce local histone hyper acetylation in regions surrounding alternative exons by inhibiting histone deacetylase 2 (HDAC2). Therefore chromatin remains hyper acetylated after the first passage of polymerase and local elongation rate is increased in later passages leading to decreased exon inclusion.

More evidence of chromatin modulating splicing activity was reported by Karen-Shaul et al. Using a plasmid reporter it was shown that strengthening 5' splice site increases inclusion of the alternative exon but also nucleosome density in the region (Keren-Shaul, Lev-Maor, and Ast). Similarly, levels of H3K36me3 can be affected by splicing activity as

illustrated by the effects of 3' splice site mutation or by inhibiting splicing using antitumor drug Spliceostatin A (Kim et al.), which binds to and inhibits components of U2 snRNP (Bonnal, Vigevani, and Valcárcel). Mechanistically, splicing can alter H3K36me3 levels by favoring the recruitment of the HYPB/Setd2 methyltransferase to the CTD of the RNA polymerase II (de Almeida et al.).

In conclusion, all the evidence points to a model where splicing impacts transcription and chromatin and both of them can feed back to splicing too. Splicing decisions appear to be guided by a very complex regulatory network, containing both feed forward and feedback circuits (de Almeida and Carmo-Fonseca).

CHAPTER 2 - OBJECTIVES

Eukaryotic genomes are packaged into chromatin, a highly regulated organized structure consisting of DNA and histone proteins. All nuclear steps of transcription take place in the context of chromatin and histone modifications have been, for a long time, associated with expression levels.

Evidence still remains correlative and recent studies showed that such relationships were actually not essential. Also, with the contribution of this thesis to the discovery that splicing mostly occurs co-transcriptionally, new steps of RNA production started being study in the context of chromatin as well. In this thesis work we aim to:

1. Describe developmentally regulated genes and their particular chromatin organizations

2. Describe co-transcriptional spliceosome assembly and to quantify the fraction of exons that are co-transcriptionally spliced

3. Identify, across multiple cell lines, exons with differentially splicing usage and new connections between chromatin and splicing in a genome-wide level

CHAPTER 3 – RESULTS

Results – Part I

Gene expression without canonical chromatin marking in developmentally regulated genes

Sílvia Pérez-Lluch[1,2,3]†, Enrique Blanco[3]†, Hagen Tilgner[1,2]†‡, Joao Curado†[1,2,4], Montserrat Corominas[3]* and Roderic Guigó[1,2]*

[1]Centre for Genomic Regulation (CRG). Doctor Aiguader 88, 08003 Barcelona, Spain.

[2]Universitat Pompeu Fabra (UPF). Doctor Aiguader 88, 08003 Barcelona, Spain.

[3]Departament de Genètica i Institut de Biomedicina (IBUB) de la Universitat de Barcelona. Diagonal 643, 08028 Barcelona, Catalonia, Spain.

[4]Graduate Program in Areas of Basic and Applied Biology, Abel Salazar Biomedical Sciences Institute, University of Porto, 4099-003 Porto, Portugal.

‡Present address: Department of Genetics, Stanford University, Stanford, California 94305, USA.

†Equal contribution

*Corresponding authors: Montserrat Corominas (mcorominas@ub.edu) and Roderic Guigó (roderic.guigo@crg.cat).

Abstract

The interplay of active and repressive histone modifications at promoter regions is assumed to play a key role in the regulation of gene expression. In contrast to this generally accepted model, we show that activation of genes regulated during metazoan development occurs in the absence of canonically active histone modifications, while strong chromatin marking is associated to transcriptional stability and tighter regulation of splicing. Consistently, promoters of developmentally regulated genes have a characteristic architecture that globally differentiates them from promoters of stably expressed genes. Our results support a model in which chromatin marking is associated to stable, tightly controlled production of RNA, while unmarked chromatin would permit rapid gene activation and de-activation during development. In these genes, Transcription Factors binding to chromatin would play a comparatively more important regulatory role.

Introduction

Epigenetic modifications in chromatin, including post-translational modifications of histone tails, are associated to the differential production of RNA that underlies cellular differentiation. Based mostly on the cumulative observation of the behavior of individual genes, an evolutionary conserved "histone code" governing differential gene expression has emerged[1]. Trimethylation of histone H3 at lysine 4 (H3K4me3) and at lysine 36 (H3K36me3), for instance, correlate with activation of transcription, whereas H3K9me3 and H3K27me3 are usually linked to transcriptional repression[2, 3]. The combinatorial behavior of histone modifications along regulatory regions—reflecting and/or influencing the specific arrangement of Transcription Factors (TFs) and other regulators—would modulate the expression levels of genes, conferring them with a unique temporal and spatial transcriptional program. Indeed, computational models have been developed that can predict gene expression from modification levels of histones in regulatory regions with great accuracy[4, 5].

A number of recent reports, however, indicate that expression of certain genes may occur in the absence of histone modifications canonically associated to active genes. The modENCODE project reported that a fraction of expressed genes lacked H3K4me3 at their annotated TSS[6]. Hödl and Basler described transcription without H3K4 methylation in the wing imaginal disc[7]. These authors found that cells that code for a non-methylable residue, instead of lysine 4, are competent to respond to developmental signaling pathways by activating target gene expression. Chen et al. also observed that pre-midblastula transition (pre-MBT) genes tend to have particularly low levels of H3K4me3, even when pre-MBT genes continue to be transcribed during the MBT[8]. More recently, Zhang et al. reported that genes located within yeast heterochromatic regions can be transcribed in absence of active histone marks[9]. Since regulated activation and de-activation of genes is particularly relevant during morphogenesis, here we investigated the relationship between histone modifications and

gene expression along development. Thus, we analyzed data produced by the modENCODE project in whole animals and tissues and characterized the fly transcriptome by RNASeq and the epigenome by ChIPSeq in two spatially well-defined and relatively homogeneous tissues: Wing and Eye-antenna imaginal discs. Finally, we carried out targeted experimental validations in isolated cells. Our analyses strongly suggest that activation of genes regulated during fly development can occur in the absence of the chromatin marks typically associated with gene activation. They also indicate that strong chromatin marking is associated not only to elevated transcriptional levels but also to stability in RNA production, both at the transcriptional and post-transcriptional levels.

Results

Expression without activation-associated histone modifications in genes regulated during development

To investigate the dynamics of chromatin marking in developmentally regulated genes, compared with that of genes stably expressed during development, we analyzed data produced within the *Drosophila melanogaster* modENCODE project[6, 10]. We specifically analyzed RNASeq data and ChIPSeq data for H3K4me3, H3K9ac, H3K4me1, H3K27ac, H3K27me3 and H3K9me3 on whole animals (Supplementary Fig. 1a). To measure transcriptional stability, we computed the coefficient of variation (cv) of gene expression over 12 developmental time points (Methods and Supplementary Fig. 1b). The cv distribution uncovers a large class of genes with low coefficient of variation, and that therefore show constant expression during development, and two other minor classes containing genes whose expression is highly variable during development—often restricted to a limited number of stages (Supplementary Fig. 1c, d). We arbitrarily selected the 1,000 genes with the highest coefficient of variation as developmentally regulated genes and matched them with the 1,000 genes with the lowest coefficient of variation as developmentally stable. For each gene, we determined the time point at which its

expression is maximal. At this time point, we did not observe differences between the expression of stable and regulated genes (Fig. 1a). Next, at the same time point, we measured the levels of the monitored histone modifications by specifically computing the height of the highest peak (measured as the log of the number of ChIPSeq reads reported by the modENCODE project) within the gene bodies. At the point of maximal gene expression, stable genes are strongly marked by histone modifications typically associated to transcription initiation: H3K4me3 and H3K9ac, and also by the enhancer modifications H3K4me1 and H3K27ac. Unexpectedly, however, developmentally regulated genes show very low levels of these modifications, which are comparable to those of silent genes (Fig. 1b, Supplementary Fig. 2). In Figure 1c we compare the pattern of H3K4me3 along fly development in *CG8636*, a gene stably expressed during development, and in *CG16733,* a gene specifically expressed in pupa. *CG8636* shows a strong H3K4me3 peak downstream from the Transcription Start Site (TSS), whereas *CG16733* lacks any marking, even at the pupa stage, where it is expressed at higher levels than *CG8636*. Lack of marking cannot be generally attributed to restricted expression patterns in regulated genes, since marking can still be detected in stable genes with restricted expression patterns (Supplementary Fig. 3). This contrasting pattern of histone marking is not only apparent when comparing genes with extreme behavior, but it is actually a distinct feature of the partition of the entire set of fly genes in two major classes according to transcriptional stability (Supplementary Fig. 4). For the repressive marks H3K27me3 and H3K9me3 we computed the average signal over the gene body and observed that regulated genes showed higher levels than stable ones, and at similar levels, even slightly higher, than silent genes (Fig. 1d). The difference of repression-associated marks between stable and silent genes is, however, small compared with the differences observed for active marks. This is partially due to a large proportion of silent genes lacking any evidence of these marks (Supplementary Fig. 5)[11].

To investigate whether lack of chromatin marking is a general feature of developmentally regulated genes across metazoans, we analyzed in a similar way *C. elegans* modENCODE data. We used RNASeq based gene expression obtained on seven time points through *C. elegans* development[12] and ChIP-chip data on the two histone modifications available for all these time points: H3K4me3 and H3K36me3. Given the smaller number of available developmental time points, we have employed a stricter criterion, and considered only the 250 genes with the lowest coefficient of variation as developmentally stable genes, and the 250 genes with the highest coefficient as developmentally regulated genes. While both, the temporal resolution and the reliability of the chromatin data obtained through ChIP-chip rather than ChIPSeq, are lower in worm than in the fly, we observed a similar trend: the expression level at the time point of maximum expression is very similar in worm regulated and stable genes (Fig. 2a), while regulated genes show much lower levels of H3K4me3 and H3K36me3, and comparable in the latter to those of silent genes (Fig. 2b).

Given that developmental chromatin maps produced in the modENCODE project are on whole organisms, it could be argued that apparent lack of chromatin marking is the consequence of the expression of developmentally regulated genes being spatially confined to specific domains, which would make it undetectable using current technologies. While, indeed, developmentally regulated genes show in general a spatially restricted pattern of expression, chromatin marking can actually be detected in stable genes that exhibit also a restricted expression pattern comparable to that in regulated genes (Supplementary Fig. 3). To further rule out this possibility we used tissue-specific RNASeq data recently released by the modENCODE fly project. Data currently available include tissues from a few developmental and adult time points[13]. Third instar larva (L3) is the time point with the larger number of tissues available: Carcass, Central nervous system, Digestive system, Fat body, Imaginal discs and Salivary glands. Collectively, they comprehend most of the

animal body at this developmental time point. Using L3 tissue-specific RNASeq data, we identified seven developmentally regulated genes (Supplementary Fig. 1c) that are expressed in all six available tissues at L3 ("Regulated broadly expressed" Fig. 3a, left panel). Conversely, we identified 130 stable genes that are specifically expressed in only one of the aforementioned tissues in L3 ("Stable tissue-specific", Fig 3a, right panel). Regulated broadly expressed genes have much higher expression levels than stable tissue specific genes when measured in the whole body (almost four-fold, Fig. 3b), as well as, in general, when measured on individual tissues (Supplementary Fig. 6). They have also higher expression levels than stable genes overall. We confirmed the expression of these genes by q-PCR (Fig. 3b). However, their H3K4me3, H3K9ac, H3K4me1, and H3K27ac levels are comparable to those in genes silent at L3, and significantly lower than in stable genes, and even than in stable tissue specific genes (Fig. 3c). We confirmed the levels of the transcription initiation marks H3K4me3 and H3K9ac by individual ChIP-qPCR (Fig. 3d).

All these results strongly suggest that activation of developmentally regulated genes occurs mostly in absence of histone modifications canonically linked to active genes. They also suggest that strong chromatin marking is associated not only to transcriptional levels, as generally accepted, but also to transcriptional stability. To explore this hypothesis, we computed the correlation in stable genes between the coefficient of variation (cv)—as a measure of transcription stability: lower cv, higher stability—and the average marking by histone modifications. For all active histone modifications, the correlation is negative and significant (as low as -0.4 for H3K4me3, Supplementary Table 2a), and comparable in magnitude to the correlation between gene expression and histone modifications (which is 0.4 for H3K4me3, when computed on averages along all developmental time points). The effect is even stronger when considering the entire set of genes, and not only the stable ones, although in this case, the results can be confounded by the more restricted

expression pattern of regulated genes (Supplementary Table 2a). The effect is not an indirect consequence of a potential correlation between transcriptional levels and transcriptional stability, since it remains when controlling for transcriptional levels using partial correlations (Supplementary Table 2b).

Gene expression without activation-associated histone modifications in imaginal discs

Data generated by the modENCODE project monitors complex systems encapsulating great cellular heterogeneity. To investigate the dynamics of chromatin marking during development in a more homogeneous cellular environment, we characterized the transcriptome by RNASeq (Supplementary Fig. 7a, b and Supplementary Table 1) and the epigenome by ChIPSeq in two *D. melanogaster* third instar larval tissues: Wing and Eye-antenna imaginal discs (WID and EID, respectively). We specifically monitored H3 and the active marks performed by modENCODE H3K4me3, H3K9ac, H3K4me1, and H3K27ac, plus the transcription elongation mark H3K36me3 (Supplementary Fig. 7c) Both, WID and EID, are tissues in early differentiation from epithelial origin, and differentially expressed genes are likely to be under developmental control. While WID and EID transcriptomes and epigenomes are very similar (Supplementary Fig. 7c-e), differentially expressed genes do exhibit functions strongly consistent with the known biology of these tissues (Supplementary Fig. 7f).

We then investigated the marking of regulated and stable genes in WID and EID. To further focus on genes under stronger regulation, we identified 55 developmentally regulated genes expressed in EID, but not in WID and 10 regulated genes, expressed in WID but not in EID. We also identified a set of 284 stable genes highly expressed both in EID and WID, as well as a set of 30 genes silent in both. (Supplementary Tables 3-8 and Methods).

We next compared marking of stable, silent, and regulated WID- and EID-specific genes (simply, WID- and EID-specific). The WID- and EID-profiles of stable genes are very similar, as there are those of silent genes (Fig. 4a), which is consistent with previous observations[11, 14]. Stable and silent genes are both characterized by higher stable nucleosome occupancy than nearby intergenic regions, but the genic nucleosome enrichment (the H3 profile) is larger for stably expressed than for silent genes. Stable genes are also strongly marked by the transcription initiation modifications H3K4me3 and H3K9ac, the elongation modification H3K36me3, and also, as observed in modENCODE, by H3K4me1 and H3K27ac. Silent genes mostly lack these histone modifications. Tissue-specific genes exhibit, however, a contrasting behavior. As expected, WID-specific genes lack H3K4me3, H3K9ac, H3K36me3, H3K4me1 or H3K27ac modifications in EID (Fig. 4b), and EID-specific genes are not marked in WID (Fig. 4c). Unexpectedly, but consistently with the behavior previously observed in modENCODE data, WID-specific genes are not marked in WID either, nor EID-specific genes in EID. Absence of active histone marking cannot be attributed to the lack of nucleosomes because a clear H3 signal is observed in these genes (Fig. 4b, c). Lack of histone marking is not due, either, to the relative low level of expression of WID- or EID-specific genes, since even for WID- or EID-specific genes with high levels of expression, comparable to those of constitutively expressed genes, there is no marking by H3 modifications associated to gene activation. This is explicitly illustrated in Figure 5 (see Supplementary Fig. 8 for more examples). WID-specific *CG4382* gene and EID-specific *CG14516* gene have similar levels of expression than the stable gene *noc*. This gene, however, is strongly marked by histone modifications in both WID and EID, while *CG4382* and *CG14516* are marked in neither. Lack of chromatin marking cannot be attributed to the restricted expression of tissue-specific genes, since the expression of *noc* is also restricted to specific regions both in WID and EID[15, 16]. H3 levels of tissue specific and stable genes are comparable and only depend weakly on the expression status of genes (Fig. 5).

Developmentally regulated genes are actively transcribed in the absence of activation-associated histone modifications

While WID and EID are relatively homogeneous tissues, they already show some cellular sub-specialization at third instar larvae. For instance, the WID-specific gene *Pdm2*, like *nubbin* (*nub*)[17], with strong temporal and spatial regulation during development, is only expressed in the wing primordium (wing pouch) at third instar larva (Fig. 6a). To unequivocally demonstrate lack of chromatin marking in developmentally regulated genes, we took advantage of the existence of the *nub*-GAL4 driver to drive expression of GFP only in the wing pouch, where *Pdm2* is expressed. Thus, we collected all cells expressing *Pdm2* and investigated chromatin marking for this gene only in the cells in which it is expressed. More specifically, dissection and dissociation of wing discs followed by cell-sorting analyses allowed the isolation of two populations of cells: the wing pouch (*nub* domain, GFP positive) and the rest of the wing (GFP negative) (Fig. 6a and Methods). By using qPCR we found that the expression of *Pdm2*, restricted to sorted GFP positive cells, is even higher than the expression of *crm*, a gene expressed at the same level throughout the WID (Fig. 6b). ChIP assays followed by qPCR on sorted cells showed that the levels of the transcription initiation mark H3K4me3, and the transcription elongation mark H3K36me3 in *Pdm2* are significantly lower than in *crm*, and comparable to those in *CG10013,* a gene silent in the whole WID (Fig. 6c).

While qPCR shows high levels of *Pdm2* in the wing pouch (Fig. 6b), RNA levels do not necessarily demonstrate active transcription, since the detected RNA molecules could have been transcribed at an earlier time point. As a measure of active gene expression we directly measured newly transcribed RNA (nascent RNA) in sorted cells. As shown in Figure 6d, *Pdm2* active transcription in GFP positive cells is as high as transcription of the control gene *crm*. *Pdm2*, therefore, is actively transcribed in the absence of active chromatin modifications.

Developmentally regulated genes exhibit characteristic promoter architecture

The striking differences in chromatin landscape between developmentally regulated and stable genes suggest that these two sets of genes may be under globally different transcriptional regulatory programs. Thus, we specifically analyzed the promoters of modENCODE stable and regulated genes. First, we found that the promoter sequence of developmentally regulated genes is significantly more conserved across the *Drosophila* genera than that of stable genes (average PhastCons[19] score of 0.27 compared to 0.17, Fig. 7a), suggesting that promoters of regulated genes are under stronger selective constraints than those of stable genes. Similar observations have been reported in mammals[20]. Conservation is particularly strong in predicted Transcription Factor (TF) Binding (TFB) motifs, with an average of 30 conserved motifs in the promoters of regulated genes compared to only 18 in stable genes (Fig. 7a). In addition, we found an enrichment of DNA Replication related Element (DRE) sequences, which are associated to disperse initiation of transcription, in promoters of stable genes. Conversely, promoters of developmentally regulated genes presented a strong overrepresentation of TATA Binding Protein (TBP) boxes, which are characteristic of tighter gene regulation[21, 22] (Fig. 7b). Consistently, we also found that promoters of stable genes overlap modENCODE High Occupancy Target (HOT) regions, associated to open chromatin and ubiquitous expression[23, 24], more often than promoters of regulated genes (67% vs. 8%). The evidence that developmentally regulated genes may be under a globally different transcriptional regulatory program than stable genes is further supported by analysis of ChIP-chip data available through the modENCODE project on 20 TFs at embryo 0-12 hours[6]. When using Principal component analysis (PCA) to classify genes expressed at this time point based on the ChIP-chip binding profiles at their promoters, a clear separation between developmentally regulated and stably expressed genes appears (Fig. 7c).

Lack of active chromatin marks is a genomic property of developmentally regulated genes independent from cell state

A number or recent reports, using complementary approaches, have described the existence of large domains underlying chromatin organization in the fly genome[25, 26]. Filion et al.[27], in particular, used integrative analysis of genome-wide maps of 53 chromatin components in the embryonic cell line Kc167 to segment the *Drosophila* genome in five main chromatin types. One of these types, labeled BLACK, covers about half of the genome, it is low in activation-associated histone modifications, and corresponds mostly to repressive chromatin. Genes in BLACK chromatin are mostly silent or expressed at very low levels in Kc167, and Filion et al. hypothesized that they were likely to be under developmental control. We have indeed found that 59% of the fly developmentally regulated genes occurred in BLACK chromatin—compared to 28% of all genes in *Drosophila* overall. In contrast, only 4% of developmentally stable genes are in BLACK chromatin (Fig. 7d).

These results suggest that developmentally regulated genes are located in repressive chromatin domains, and that lack of histone modifications is not a transient genome property dependent on a cell type since it is observed both in developmental tissues and in cell lines.

Strong chromatin marking in stably expressed genes is associated to tighter regulation of alternative splicing

Beyond its role in primary RNA production, chromatin structure has also been recently implicated in subsequent steps of RNA processing. A number of studies have uncovered a relationship between nucleosome occupancy and exon-intron structure[28, 29] and between specific histone modifications and alternative splicing[30-32]. We specifically investigated whether strongly positioned nucleosomes directly associate with exon inclusion. To isolate the effects of nucleosome occupancy on exon

inclusion from those of gene expression we computed the inclusion level of individual fly internal exons in WID and EID in a manner that is independent from overall gene expression (Methods). Then, within each tissue we selected highly included exons (inclusion level greater than 0.9) and lowly included exons (inclusion level lower than 0.1, Supplementary Tables 1, 9 and 10). Occupancy profiles at the exons' acceptor site indicate that in both WID and EID, highly included exons are characterized by higher H3 occupancy when compared to lowly included ones (Fig. 8a, b), as it has been previously reported in mammals[29]. To characterize the relationship between nucleosome occupancy and exon inclusion we computed the correlation between these two variables on a running 250bp window centered at each nucleotide within the region -1,000 to +1,000 from the acceptor site (Methods, Fig. 8c, d). The behavior of the correlation between H3 and exon inclusion is very similar for EID and WID: it significantly peaks very close to the acceptor site and essentially vanishes beyond 500 nucleotides from the acceptor site. This shows that the association of nucleosomes and exon inclusion is local, consistent across tissues and not negligible.

We speculated, thus, that strong chromatin marking might not be only associated to more stable RNA production, but also to a tighter regulation of alternative splicing. To measure alternative splicing complexity, we computed the Shannon's entropy on the relative abundance of a gene's alternative splicing isoforms (Methods). The splicing entropy grows with the number of isoforms and with the evenness of their relative abundances. The entropy is zero when there is only one isoform being expressed (which would correspond to tight regulation of isoform expression), and it reaches its maximum when all isoforms are equally expressed (which would correspond to lack of splicing regulation and stochastic production of alternative splicing isoforms). Based on transcript quantifications produced by the modENCODE project for the fly, and computed by us for the worm (see Methods), we have calculated the splicing entropy of each gene at the developmental time point in which its

expression is at its maximum. As hypothesized, splicing entropy is lower for strongly marked stable genes than for unmarked developmentally regulated genes in both fly and worm for any number of annotated isoforms (Fig. 8e). Further supporting tighter regulation of splicing, we have also found that the major isoform captures a larger fraction of the total transcriptional output of the gene in stable than in regulated genes (Fig. 8f).

Discussion

Cell type specific transcriptional regulation is crucial to maintain cell identity throughout the lifetime of an organism, yet it must be flexible enough to allow for responses to endogenous and exogenous stimuli. This regulation is mediated by specific molecular factors (e.g. cell type specific transcription factors, and chromatin modifications), as well as by the topological organization of the genome. In particular, modifications occurring on DNA and on histone proteins regulate gene expression by establishing and maintaining specific chromatin states[33, 34]. It has become widely accepted that the association of certain modifications with transcriptional activation or repression is a general phenomenon. Nevertheless, expression of genes in the absence of chromatin marks has already been reported[6, 7, 9]. Here we found that transcriptional activation in the absence of most canonically active chromatin marks is actually a characteristic feature of genes that are regulated during development. We also found that strong chromatin marking appears to confer transcription stability.

Analyses of tissue-specific gene expression data, as well as our targeted validation experiments, demonstrate that our observations do not arise from the expression of developmental-specific genes being lower or confined to small cell populations, from limited detection sensitivity, and/or from persistence in the cell of RNA molecules transcribed at some earlier standpoint. Our analyses further indicate that lack of chromatin

marking is not a transient genome state, but a constitutive property of these genes. Distinct dynamic properties of promoters differentially used during the zygotic genome activation have already been described in *Drosophila*, where pre-midblastula transition (pre-MBT) genes tend to have low levels of H3K4me3, even when pre-MBT genes continue to be transcribed during the MBT[8]. Of note, we found 25 out of 65 developmentally regulated genes expressed at early embryonic stages among the 117 pre-MBT genes reported by Zeitlinger and colleagues[8] in contrast to only one out of the 1,000 stable genes.

We also found that a highly structured and strongly marked chromatin state leads to tightly controlled RNA production, not only at transcription but also at splicing level. We indeed observed higher stochasticity in the production of alternative splice forms in unmarked developmentally regulated genes than in marked stably expressed ones. Tighter regulation of splicing by chromatin marking is consistent with earlier observations[35] of simultaneous enrichment in the expression of chromatin modifying enzymes and splicing factors in cell-enriched testis. It is also consistent with the higher levels of H3K36me3 found by de Almeida et al.[30] in constitutive exons compared to alternative exons in mammalian genomes (a finding that we have also replicated in the fly genome, data not shown).

Overall, our results lead us to hypothesize that the relative contribution of Transcription Factors and Histone Modifications defines two major broad transcriptional regulatory programs. In stable that are constitutively expressed, strong chromatin marking leads to transcriptional stability and tightly controlled RNA production. In these genes, regulation by Transcription Factors would play a comparatively smaller role. In contrast, developmentally regulated genes that need to be rapidly activated and de-activated are characterized by an unmarked chromatin state. In these genes, Transcription Factors binding to

chromatin would play the predominant regulatory role. It is unclear whether this model of transcriptional regulation can be generalized to other metazoans. While detailed transcriptional and epigenetic maps are being produced in an increasing number of cell lines and tissues (both healthy and diseased), developmental maps are still sparse in mammalian species. Exhaustive monitoring through a much larger variety of conditions, differentiation states and developmental stages is required to fully understand the layer of epigenetic regulation that mediates between genome sequence and RNA production.

Acknowledgements

We thank D. Gonzalez-Knowles, A. Breschi and M. Melé, for help with data analysis. We thank the modENCODE consortium for granting open access of these resources to the scientific community. We also thank the Ultrasequencing Unit of the CRG (Barcelona, Spain), for sample processing. This work was performed under the financial support of the Spanish MICINN with grants BIO2011-26205 and of the ERC/European Community PF7 with grant 294653 RNA-MAPS to R. G., and grants CSD2007-00008 and BFU2102-36888 to M. C. J. C. is supported by grant SFRH/BD/33535/2008 from the Portuguese Foundation to Science and Technology.

Contributions

S. P-Ll. performed the experiments. E. B., H. T. and J. C. performed the computational analysis; M. C. and R. G designed the analysis and wrote the paper; all authors discussed the data.

Competing financial interests

The authors declare no competing financial interests.

Corresponding author

Correspondence to: Roderic Guigó (roderic.guigo@crg.cat) and Montserrat Corominas (mcorominas@ub.edu).

Online Methods

***Drosophila* strains**

The strains used were: *Canton S* as a wild- type and *nub*-GAL4/+; UAS-GFP/+. Flies were kept on standard media at 25ºC.

Tissue disaggregation and cell sorting

Wing imaginal discs (WID) from *nub*-GAL4/+; UAS-GFP/+ flies were dissected in PBS and incubated for 1h in a 10x trypsin solution (Sigma T4174) at room temperature in a rotating wheel. Cells were vigorously

pipetted and kept on ice in Schneider's insect medium. To discard dead cells, DAPI was added to the sample at 1 mg/mL final concentration. Cells were sorted in a FACSAria (BD) with the 85 mm nozzle. We were able to recover around $2.5 \cdot 10^6$ GFP negative and $2 \cdot 10^6$ GFP positive cells from 400 WIDs. An independent sorting experiment was done per each replicate, both for ChIPs and gene expression analyses.

RNA extraction, retrotranscription and Real-Time PCR

As starting material, 120 WID and 250 eye-antenna imaginal discs (EID) were used for RNASeq. For *Pdm2* gene expression analysis, WIDs from 400 *nub*-GAL4/+; UAS-GFP/+ flies were disaggregated. RNA from sorted cells was extracted with ZR-RNA MicroPrep Kit from Zymo Research. For L3-specific genes expression, 5 third instar larvae were frozen and RNA was extracted with Quick-RNA MiniPrep Kit, from Zymo Research. Retrotranscriptions and qPCRs were performed as described previously[11]. For quantification of RNA amounts, standard curves of each pair of primers were performed and the efficiency of amplification was calculated. The Cts obtained from the qPCR were corrected according to the amplification efficiency of the primers. Primers used for Real-Time PCR are listed below.

***In situ* hybridizations**

In situ hybridizations using digoxigenin labelled riboprobes were carried out according to standard protocols. The probe for *Pdm2 in situ* was PCR amplified using primers listed below and cloned into a pBSK+/- vector at *EcoR*I restriction site. Riboprobes were synthesized using T7 polymerase. WID and EID were analyzed with a Leica DMLB microscope.

Chromatin immunoprecipitation

Third instar larva WID or EID isolated from *Canton S* flies were fixed, pooled in 700 mL and processed as described[11]. From 300 to 600 imaginal discs were used in these experiments. Trypsin treated cells from GFP transgenic flies were fixed after sorting for 10 minutes at room temperature and sonicated in a Diagenode Bioruptor for 15 minutes at

high power in lysis buffer (1% SDS, 10 mM Tris HCl ph 8.0 and 2mM EDTA). Immunoprecipitations were performed in RIPA buffer. For L3 ChIPs and Imaginal Discs ChIPSeq experiments we used 1 mg of the corresponding antibody. For ChIPs in sorted cells we used 0.45 mg of anti-H3K4me3, 0.3 mg of anti-H3K36me3, 0.33 mg of anti-H3K27ac and 1 mg of anti-H3K27me3. For L3 time-specific ChIPs, 5 *Canton S* wall-wandering third instar larvae were disrupted and fixed 10 minutes at room temperature. Fixed larvae were sonicated in a Diagenode Bioruptor for 15 minutes at high power in lysis buffer. Immunocomplexes were recovered with Invitrogen ProteinA magnetic beads for 2h. The beads were washed three times in RIPA or IP buffer for 10 minutes, once in LiCl buffer and twice in TE[11].

The primers used for Real-Time PCR are listed below. The antibodies used for ChIP were: H3 (Abcam/ab1791); H3K4me3 (Abcam/ab8580) (Millipore-Upstate/07-473), H3K9ac (Abcam/ab4441), H3K36me3 (Abcam/ab9050), H3K4me1 (Diagenode/CS-037-100) and H3K27ac (Abcam/ab4729)..

Nascent RNA

For Nascent RNA assays, 400 WIDs *nub*-GAL4/+; *UAS*-GFP/+ were dissected and disaggregated as described above. Click-IT® Nascent RNA Capture Kit from Molecular Probes (cat. number C10635) was used according to the manufacturer's instructions. Briefly, disaggregated cells were incubated with 0.5 mM 5-ethynil uridine (EU) in Schneider's Insect Medium for 1 h at room temperature. Total RNA was extracted and biontinylated with 0.25 mM biotin-azide for 30 minutes at room temperature. Biotinylated RNA was precipitated over night at -80ºC and purified with Streptavidin conjugated beads for 30 minutes at room temperature. Nascent RNA was eluted in 0.1 % SDS 5 minutes at 99ºC and retrotranscription was carried out as described above. Four biological replicates were performed. Primers used for Real-Time PCR are listed below.

Primers for cloning and qPCR

Name of the primer	Sequence (5'-3')	Experiment
Pdm2_Fw	cgggctgcaggaattcGGTGGAGTGTTCAACATC	probes cloning
Pdm2_Rv	gcttgatatcgaattcATGTGGAATGTTGTGGAAG	probes cloning
Pdm2_TSS_Fw	GATCACCTCACCTCACCTC	gene expression and ChIPs
Pdm2_TSS_Rv	CAGCTTCACGCCCTTTCTG	gene expression and ChIPs
Pdm2_3'_Fw	AACTAGGTGATAGCCAGTCG	ChIPs
Pdm2_3'_Rv	TATGCATACAGATACGAATAGG	ChIPs
CG10013_TSS_Fw	AGCAAGAACTGGATCAATGTC	gene expression and ChIPs
CG10013_TSS_Rv	AGTAGGGGTTGGAAAAGATC	gene expression and ChIPs
crm_TSS_Fw	GACAATTGTGCATGGTCACC	gene expression and ChIPs
crm_TSS_Rv	GTTGTTAACTCGATAGCTAAGC	gene expression and ChIPs
crm_3'_Fw	CACCCTTAGCCACTATCAAC	ChIPs
crm_3'_Rv	AGGCGGGAGAATCTCTATAC	ChIPs
AbdB_Fw	AGACATGCGAGTAGAAAGCC	ChIPs
AbdB_Rv	TTTCGTGCTGTCTTTTGTGC	ChIPs
RpL32_Fw	CACTCACCGACAGCTTAGC	ChIPs
RpL32_Rv	CCAAGATCGTGAAGAAGCG	ChIPs
Bmcp_Fw	CTGGGACGTCGGCTATATG	ChIPs
Bmcp_Rv	AAGGTGCCTGCAATGAATAG	ChIPs
Bmcp_RNA_Fw	TCGATTCCGGACATCAACAG	gene expression
Bmcp_RNA_Rv	CGCCAATCCTTCACTTCAC	gene expression
Lsp1alpha_Fw	ACCACAAGACCCACGACATC	gene expression and ChIPs
Lsp1alpha_Rv	CTTCATCAGGAACGCCTTG	gene expression and ChIPs
Lsp1beta_Fw	TATCCGGAGCAGTTGAGTG	gene expression and ChIPs
Lsp1beta_Rv	TTCATGTTGACGGATTCCTG	gene expression and ChIPs
Lsp1gamma_Fw	TAGCAGCGAGAGGACCAAG	gene expression and ChIPs
Lsp1gamma_Rv	GCTAAAGGCAGTCACACAG	gene expression and ChIPs
Lcp1_Fw	AAGTCAGCCAATATGTTCAAG	gene expression and ChIPs
Lcp1_Rv	TAGGCATTCGTTAGGGATCG	ChIPs
Lcp1_RNA_Rv	GCCCAAAACTGCGCAGATC	gene expression
Lcp2_Fw	CGATGATGTACACGCTGATG	gene expression and ChIPs
Lcp2_Rv	TCGATTCCGTTTGAGGTGTG	gene expression and ChIPs
Lcp4_Fw	AAGGAGCTGGTCAACGATG	gene expression and ChIPs
Lcp4_Rv	TGGACATCTCCGGTAGCAG	gene expression and ChIPs
CG15353_Fw	TGAAGCTGGTGAGTGTTTGC	ChIPs
CG15353_Rv	AACAAAGGGCATTCGTTCAG	ChIPs
CG15353_RNA_Fw	AACTCAATATCCGGATTCACC	gene expression
CG15353_RNA_Rv	ACTGCTTGTCCATGTCCATG	gene expression
CG5367_Fw	TTCCTCGAACGCCTTGTAAC	gene expression and ChIPs
CG5367_Rv	AATCGCAAATATCTTCGCACC	gene expression and ChIPs

Solexa/Illumina sequencing

Solexa/Illumina sequencing was carried out at the Ultrasequencing Unit of the Centre for Genomic Regulation (CRG, Barcelona, Spain. All protocols for Solexa/Illumina ChIPSeq and for RNASeq analysis were carried out following the manufacturer's protocol. For ChIPSeq, 10 ng of each sample were used and fragments between 300 and 350 bp were size selected before sequencing. For RNASeq, 5 mg of total RNA were used to sequence.

Drosophila melanogaster genome and annotation

We used the FlyBase[13] annotation release 5.12 for the genome version dm3.

RNASeq and ChIPSeq read mapping

Reads of 36 and 40 bp obtained from single-end RNASeq and ChIPSeq sequencing from WID and EID-cells were aligned using GEM[36] allowing up to two mismatches to the *D. melanogaster* genome (version dm3) and, for RNA, to all possible junctions of 5'-3'-ordered exon pairs occurring within the same annotated gene. ChIPSeq and RNASeq raw data and profiles of read counts were deposited in the NCBI-GEO repository under the accession number GSE56551.

Gene and transcript expression analysis

Reads mapping uniquely to the genome were used to quantify genes and transcripts separately in each tissue using the FluxCapacitor[37]. Expression levels are given in Reads Per Kilobase per Million mapped reads (RPKM). Linear regression analysis between log transformed WID and EID RPKMs gave a highly significant slope and intercept. Thus, we identified 628 genes at least one unit above the linear regression line (differentially expressed genes in EID) and 184 genes at least one unit below (differentially expressed genes in WID). To build our collection of regulated tissue-specific genes from each differentially expressed gene set, we required cv >= 1.2 and at least 1.5 RPKMs in one tissue and less than 0.1 RPKM in the other one (55 EID-specific genes and 10 WID-specific genes, respectively, resulted from this criterion). Finally, those genes with cv < 1.2 that are expressed in both tissues (> 2.3 RPKMs) with a difference in expression of less than 20% were selected as stable expressed in the two tissues (284 genes) and the genes whose expression in both tissues is 0 RPKMs were considered to be silent (30 genes).

ChIPSeq analyses

ChIPSeq reads for H3, H3K4me3, H3K9ac, H3K36me3, H3K4me1 and H3K27ac were extended to the full average fragment length in the corresponding experiment. For each position in the genome the number of extended ChIPSeq reads overlapping this position was recorded. Each sample was normalized by the total number of sequenced reads and the average fragment length. The genome-wide correlation between WID and EID samples was computed using the UCSC Table browser on windows of 1,000 nucleotides[38]. To compute the correlation between ChIPSeq samples and RNASeq expression data, we assigned to each gene the highest peak of the corresponding ChIP signal within the body gene and correlate this value to the expression of the gene. To produce the graphical distribution of reads for each sample around a particular site (Transcription Start Sites, TSS, polyAdenylation Sites, pAS and splice Acceptor Sites, AS), we calculated the weighted number of reads on each position from -500 bp to +500bp of each TSS, pA and AS, according to FlyBase. To graphically represent an idealized gene, we normalized the location of the reads within the gene using a window of 100 units, and calculated the mean at each point. We extended this representation 500 bps upstream and downstream of the gene. To compare WID and EID samples, we calculated the weighted number of reads on each position in the normalized ChIPSeq profiles.

modENCODE analyses

Stable and developmentally regulated genes in *D. melanogaster*

To define the transcriptional stability of genes, we calculated the coefficient of variation of gene expression, as reported by the modENCODE consortium[10], for each protein-coding gene that has detectable expression in 12 selected time points (Supplementary Fig. 1a). From the full ranking of 13,635 genes, we defined the bottom 1,000 genes with lowest variation of expression during development as stable, and the top 1,000 genes with highest variation as developmentally regulated genes. In addition, at each time point we selected the same number of silent genes than

developmentally regulated genes expressed at that time point, for a total of 1,000 silent genes. For these genes, we measured the strength of the highest peak (measured as the log of the number of reads reported by modENCODE) within the gene body at the time point in which its expression is maximum for H3K4me3, H3K9ac, H3K4me1, H3K27ac, and the average signal within the gene body for H3K27me3 and H3K9me3 modENCODE ChIPSeq profiles (NCBI GEO accession: GSE16013). Due to data issues with ChIPSeq for three samples: H3K9ac (Adult male) and H3K9me3 (L3 and Adult male), we used ChIP-chip data in these cases instead. The Wilcoxon test (two-sided) was used to evaluate the statistical significance of the difference between ChIP values for stable and developmentally regulated genes on each sample. To build the subsets of low, medium and high regulated genes, we ranked the top 1,000 regulated genes by their expression (in the time point of maximum expression) and we classified them into three groups of the same number of genes.

Stable and developmentally regulated genes in *C. elegans*

We estimated H3K4me3 and H3K36me3 levels in 7 developmental stages (Early Embryo, Late Embryo, Larvae L1, L2, L3, L4 and Young Adult) from array signal files in Gerstein et al[24]. To define developmentally stable and regulated genes, we also used the same procedure as in fly. To obtain gene and transcript quantifications, we mapped the RNASeq reads from modENCODE *C. elegans*[24] to the Wbcel215.68 version of the genome using GEM[36], and used the FluxCapacitor[37] to produce the quantifications.

L3-specific genes analysis

To compare the expression and histone modification marking levels in regulated broadly expressed and stable tissue-specific genes we used anatomy RNASeq data from modENCODE consortium available in Flybase[13]. We used the gene sets previously defined for modENCODE analysis to create new subgroups of genes:

- Stable: the 1,000 genes with the lowest coefficient of variation of gene expression across modENCODE time points
- Silent: genes identified as silent in L3 stage (RPKM=0)
- Regulated broadly expressed at L3: developmentally regulated genes that are detected in L3 whole body data, and that are furthermore expressed with at least 1 RPKM in each of the 6 tissues with L3 tissue-RNASeq available
- Stable tissue-specific at L3: from the set of extended stably expressed genes (P1 in Supplementary Fig 4) we selected the genes that, using L3 tissue-RNASeq, are detected as expressed with at least 10 RPKM in 1 of the tissues and not higher than 1 in all the other remaining tissues. We identified 26 carcass-specific genes, 8 central nervous system-specific genes, 36 digestive-specific genes, 21 fat body-specific genes, 36 imaginal disc-specific genes and 4 salivary glands-specific genes.

The expression and histone modification levels were calculated using L3 data from modENCODE following the methodology of the previous analysis.

Promoter analyses

To measure the conservation of the promoters of regulated and stable genes across 12 Drosophilids, we computed the average of the UCSC PhastCons multiz15way track[38] along the promoter sequences of each gene set (promoter length: 200 bp). To characterize the promoters of regulated and stable genes, we used the MatScan program[39] with the full collection of 827 predictive matrices available in Jaspar and Transfac[40, 41]. From each initial pool of predictions, we removed those binding sites within genome regions in the UCSC genome browser that presented on average a probability lower than 0.95 to be conserved across the 12 flies PhastCons multiz15way alignments[19]. The Wilcoxon test (one-sided) was used to evaluate the statistical significance of the difference for stable and developmentally regulated gene sets on each comparison (PhastCons scores and number of conserved sites). For the identification of

focused/dispersed initiation sites[8, 22], we searched for putative binding sites of TBP, GAGA, Zelda and DRE in the promoter sequence of the top 1,000 stable and the top 1,000 regulated genes (promoter length: 100 bp). We selected TBP as a marker of focused initiation and DRE as a representative of dispersed initiation. GAGA and Zelda are not characteristic of a particular transcriptional model. The weight matrices for TBP and GAGA are from Jaspar[40] and for Zelda and DRE are from Fly Factor Survey[42].

We performed principal components analysis (PCA) based on the ChIPSeq levels of 20 Transcription Factors in Embryos at 0-12h in the promoter regions of genes with expression above 10 as measured by tilling arrays at this time point[6].

Splicing entropy

For each gene, we computed the Shannon's entropy (or diversity index) based on the relative frequencies of the gene's annotated isoforms in a given cell line. Let *g* be a gene with *n* annotated isoforms with relative frequencies $p_1, \ldots, p_n$, in a given condition, the entropy of *g*, *H(g)*, is computed as

$$H(g) = -\sum_{i=1}^{n} p_i \ln p_i$$

H(g) growths with the number of annotated isoforms and with the evenness of their frequencies. *H(g)* is zero when there is only one expressed isoform, and it is maximum when all isoforms are equally expressed. The boxplots in Fig. 8c, d display the distribution of *H(g)*, separately for genes with different number of isoforms.

References

1. Li, B., Carey, M. & Workman, J.L. The role of chromatin during transcription. *Cell* **128**, 707-19 (2007).
2. Black, J.C., Van Rechem, C. & Whetstine, J.R. Histone lysine methylation dynamics: establishment, regulation, and biological impact. *Mol Cell* **48**, 491-507 (2012).
3. Wagner, E.J. & Carpenter, P.B. Understanding the language of Lys36 methylation at histone H3. *Nat Rev Mol Cell Biol* **13**, 115-26 (2012).
4. Dong, X. et al. Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol* **13**, R53 (2012).
5. Karlic, R., Chung, H.R., Lasserre, J., Vlahovicek, K. & Vingron, M. Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci U S A* **107**, 2926-31 (2010).
6. Negre, N. et al. A cis-regulatory map of the Drosophila genome. *Nature* **471**, 527-31 (2011).
7. Hodl, M. & Basler, K. Transcription in the absence of histone H3.2 and H3K4 methylation. *Curr Biol* **22**, 2253-7 (2012).
8. Chen, K. et al. A global change in RNA polymerase II pausing during the Drosophila midblastula transition. *Elife* **2**, e00861 (2013).
9. Zhang, H., Gao, L., Anandhakumar, J. & Gross, D.S. Uncoupling transcription from covalent histone modification. *PLoS Genet* **10**, e1004202 (2014).
10. Graveley, B.R. et al. The developmental transcriptome of Drosophila melanogaster. *Nature* **471**, 473-9 (2011).
11. Perez-Lluch, S. et al. Genome-wide chromatin occupancy analysis reveals a role for ASH2 in transcriptional pausing. *Nucleic Acids Res* **39**, 4628-39 (2011).
12. Spencer, W.C. et al. A spatial and temporal map of C. elegans gene expression. *Genome Res* **21**, 325-41 (2011).
13. Tweedie, S. et al. FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res* **37**, D555-9 (2009).
14. Barski, A. et al. High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823-37 (2007).

15. Luque, C.M. & Milan, M. Growth control in the proliferative region of the Drosophila eye-head primordium: the elbow-noc gene complex. *Dev Biol* **301**, 327-39 (2007).
16. Weihe, U., Dorfman, R., Wernet, M.F., Cohen, S.M. & Milan, M. Proximodistal subdivision of Drosophila legs and wings: the elbow-no ocelli gene complex. *Development* **131**, 767-74 (2004).
17. Ng, M., Diaz-Benjumea, F.J. & Cohen, S.M. Nubbin encodes a POU-domain protein required for proximal-distal patterning in the Drosophila wing. *Development* **121**, 589-99 (1995).
18. Adelman, K. & Lis, J.T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat Rev Genet* **13**, 720-31 (2012).
19. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-50 (2005).
20. Rodelsperger, C. et al. Short ultraconserved promoter regions delineate a class of preferentially expressed alternatively spliced transcripts. *Genomics* **94**, 308-16 (2009).
21. Ohler, U., Liao, G.C., Niemann, H. & Rubin, G.M. Computational analysis of core promoters in the Drosophila genome. *Genome Biol* **3**, RESEARCH0087 (2002).
22. Juven-Gershon, T. & Kadonaga, J.T. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev Biol* **339**, 225-9 (2010).
23. Farley, E. & Levine, M. HOT DNAs: a novel class of developmental enhancers. *Genes Dev* **26**, 873-6 (2012).
24. Gerstein, M.B. et al. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* **330**, 1775-87 (2010).
25. Hou, C., Li, L., Qin, Z.S. & Corces, V.G. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Mol Cell* **48**, 471-84 (2012).
26. Roy, S. et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330**, 1787-97 (2010).
27. Filion, G.J. et al. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell* **143**, 212-24 (2010).

28. Schwartz, S., Meshorer, E. & Ast, G. Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol* **16**, 990-5 (2009).
29. Tilgner, H. et al. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol* **16**, 996-1001 (2009).
30. de Almeida, S.F. et al. Splicing enhances recruitment of methyltransferase HYPB/Setd2 and methylation of histone H3 Lys36. *Nat Struct Mol Biol* **18**, 977-83 (2011).
31. Luco, R.F. et al. Regulation of alternative splicing by histone modifications. *Science* **327**, 996-1000 (2010).
32. Sims, R.J., 3rd et al. Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. *Mol Cell* **28**, 665-76 (2007).
33. Delest, A., Sexton, T. & Cavalli, G. Polycomb: a paradigm for genome organization from one to three dimensions. *Curr Opin Cell Biol* **24**, 405-14 (2012).
34. Espada, J. & Esteller, M. DNA methylation and the functional organization of the nuclear compartment. *Semin Cell Dev Biol* **21**, 238-46 (2010).
35. Gan, Q. et al. Dynamic regulation of alternative splicing and chromatin structure in Drosophila gonads revealed by RNA-seq. *Cell Res* **20**, 763-83 (2010).
36. Derrien, T. et al. Fast computation and applications of genome mappability. *PLoS One* **7**, e30377 (2012).
37. Montgomery, S.B. et al. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773-7 (2010).
38. Karolchik, D. et al. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res* **36**, D773-9 (2008).
39. Blanco, E., Messeguer, X., Smith, T.F. & Guigo, R. Transcription factor map alignment of promoter regions. *PLoS Comput Biol* **2**, e49 (2006).
40. Portales-Casamar, E. et al. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res* **38**, D105-10 (2010).

41. Wingender, E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief Bioinform* **9**, 326-32 (2008).
42. Zhu, L.J. et al. FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Res* **39**, D111-7 (2011).

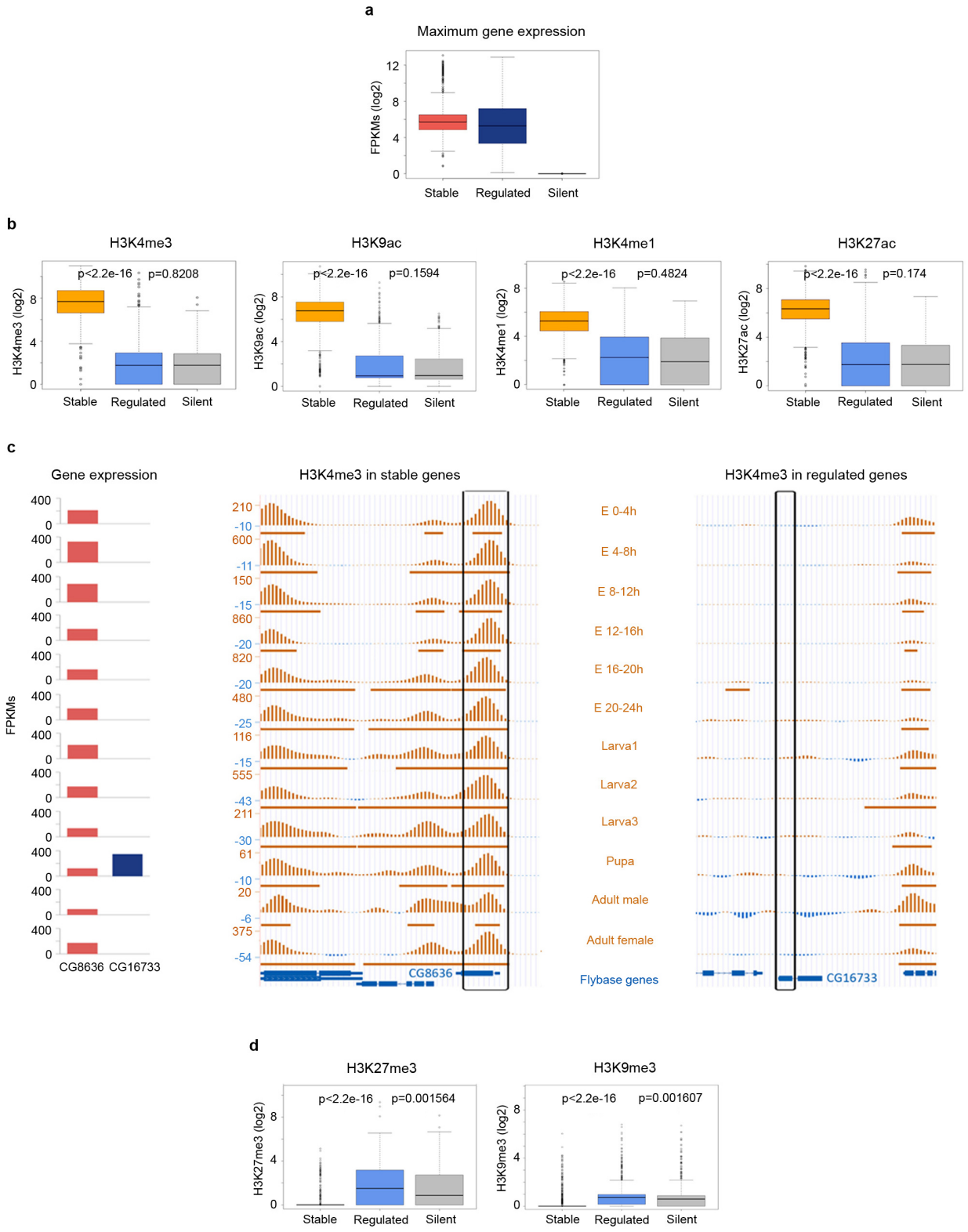

a
Maximum gene expression
FPKMs (log2)
Stable
Regulated
Silent
b
H3K4me3
H3K9ac
H3K4me1
H3K27ac
p<2.2e-16
p=0.8208
p=0.1594
p=0.4824
p=0.174
c
Gene expression
FPKMs
CG8636
CG16733
H3K4me3 in stable genes
H3K4me3 in regulated genes
E 0-4h
E 4-8h
E 8-12h
E 12-16h
E 16-20h
E 20-24h
Larva1
Larva2
Larva3
Pupa
Adult male
Adult female
Flybase genes
d
H3K27me3
H3K9me3
p<2.2e-16
p=0.001564
p=0.001607

Figure 1

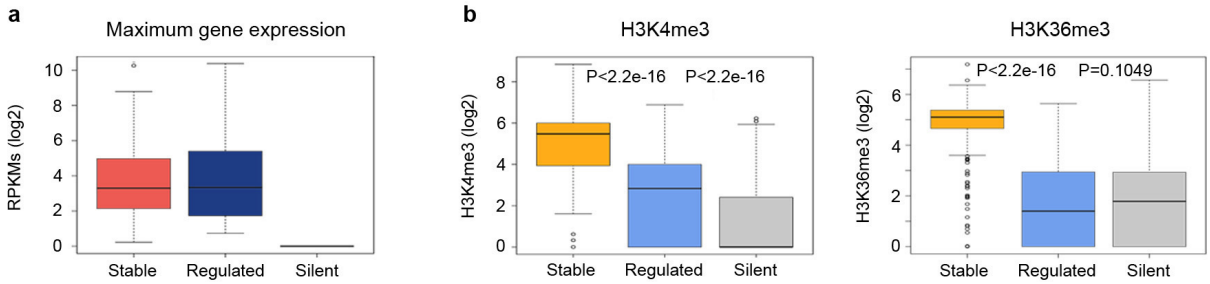

a
Maximum gene expression
RPKMs (log2)
10
8
6
4
2
0
Stable
Regulated
Silent
b
H3K4me3
H3K4me3 (log2)
P<2.2e-16
P<2.2e-16
8
6
4
2
0
Stable
Regulated
Silent
H3K36me3
H3K36me3 (log2)
P<2.2e-16
P=0.1049
6
4
2
0
Stable
Regulated
Silent

Figure 2

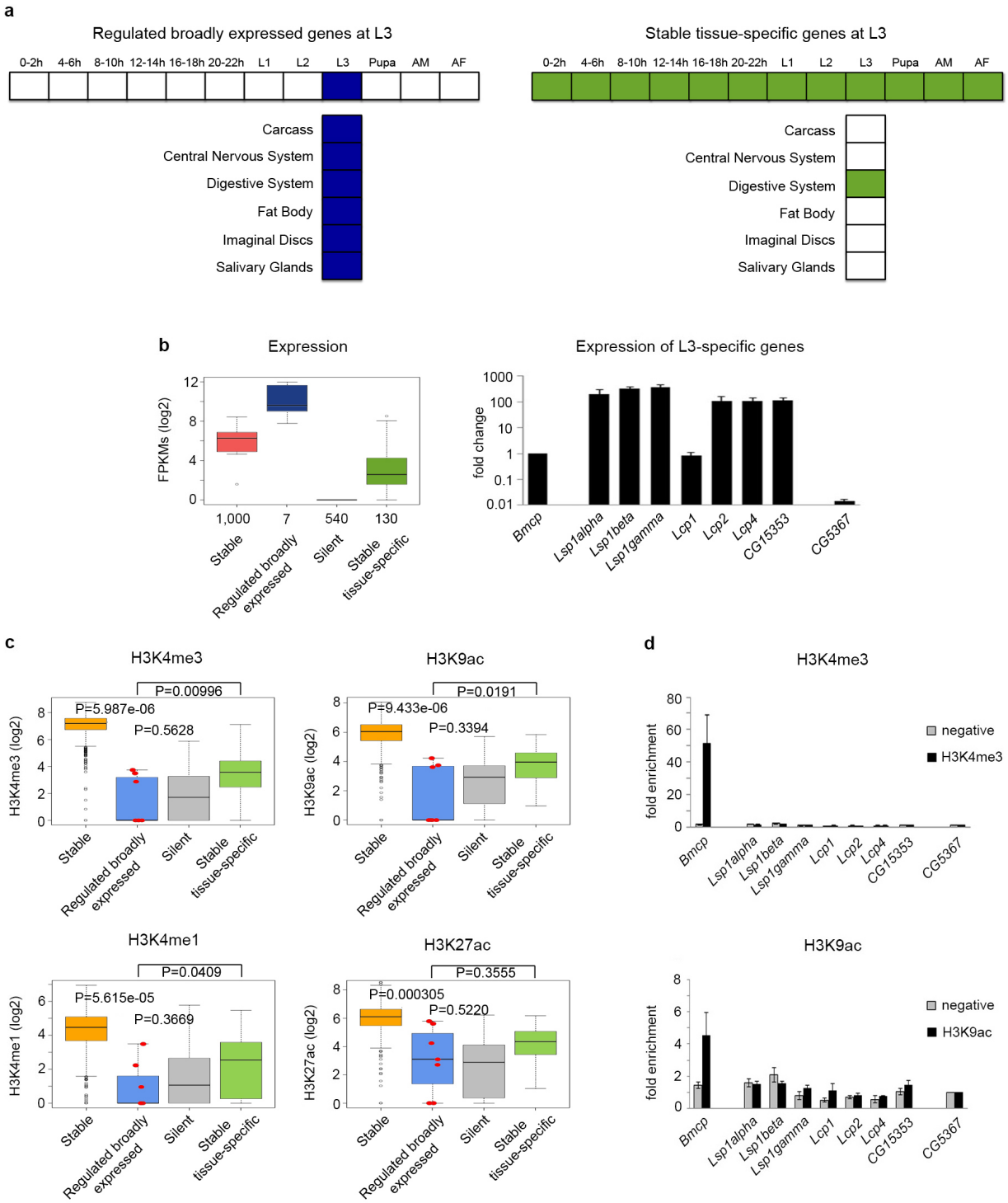

a
Regulated broadly expressed genes at L3
Stable tissue-specific genes at L3
0-2h 4-6h 8-10h 12-14h 16-18h 20-22h L1 L2 L3 Pupa AM AF
Carcass
Central Nervous System
Digestive System
Fat Body
Imaginal Discs
Salivary Glands
b
Expression
FPKMs (log2)
1,000
7
540
130
Stable
Regulated broadly expressed
Silent
Stable tissue-specific
Expression of L3-specific genes
fold change
Bmcp
Lsp1alpha
Lsp1beta
Lsp1gamma
Lcp1
Lcp2
Lcp4
CG15353
CG5367
c
H3K4me3
P=0.00996
P=5.987e-06
P=0.5628
H3K9ac
P=0.0191
P=9.433e-06
P=0.3394
H3K4me1
P=0.0409
P=5.615e-05
P=0.3669
H3K27ac
P=0.3555
P=0.000305
P=0.5220
d
H3K4me3
fold enrichment
negative
H3K4me3
H3K9ac
negative
H3K9ac

Figure 3

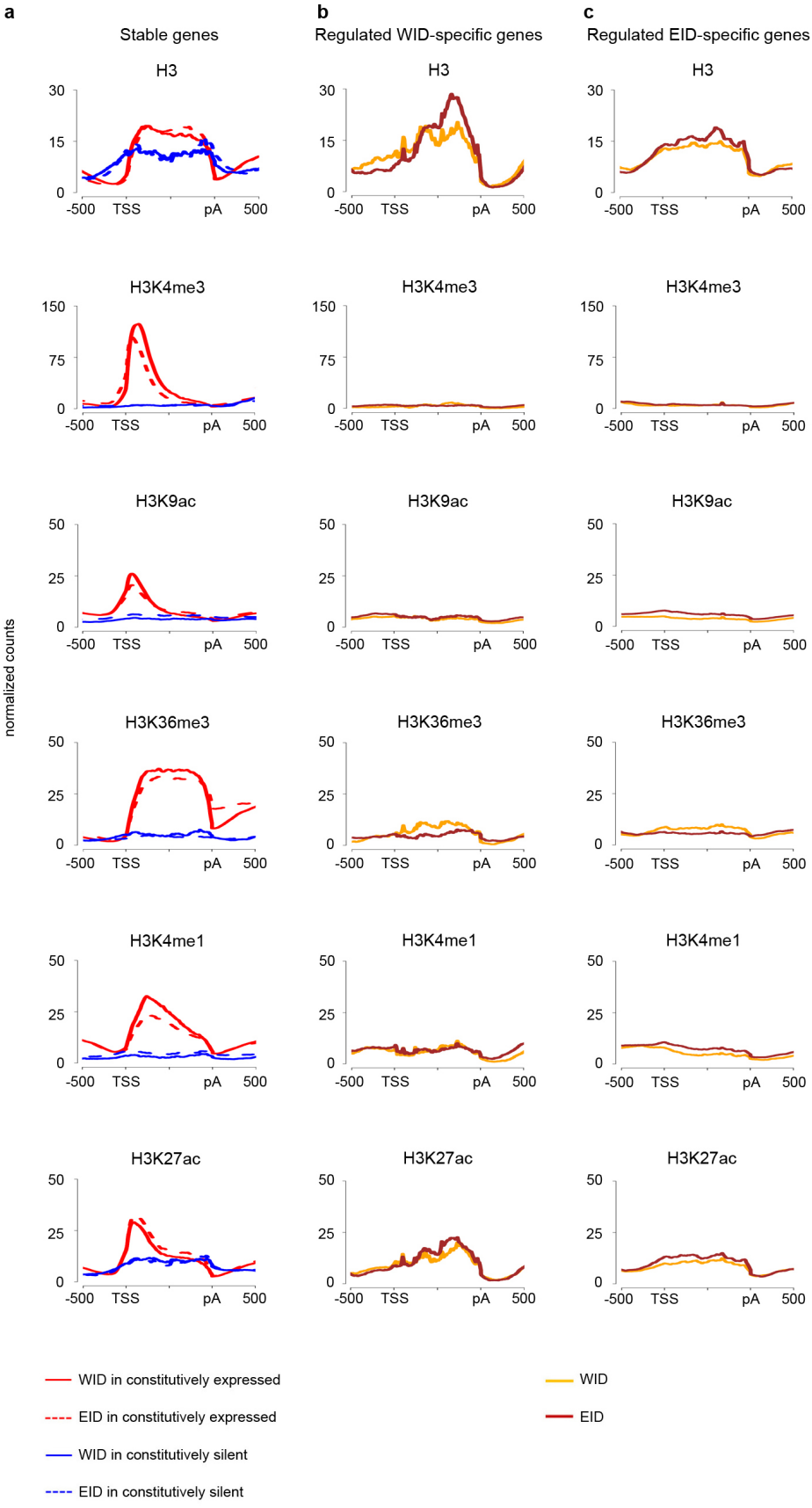

a
b
c
Stable genes
Regulated WID-specific genes
Regulated EID-specific genes
H3
H3K4me3
H3K9ac
H3K36me3
H3K4me1
H3K27ac
normalized counts
-500
TSS
pA
500
WID in constitutively expressed
EID in constitutively expressed
WID in constitutively silent
EID in constitutively silent
WID
EID

Figure 4

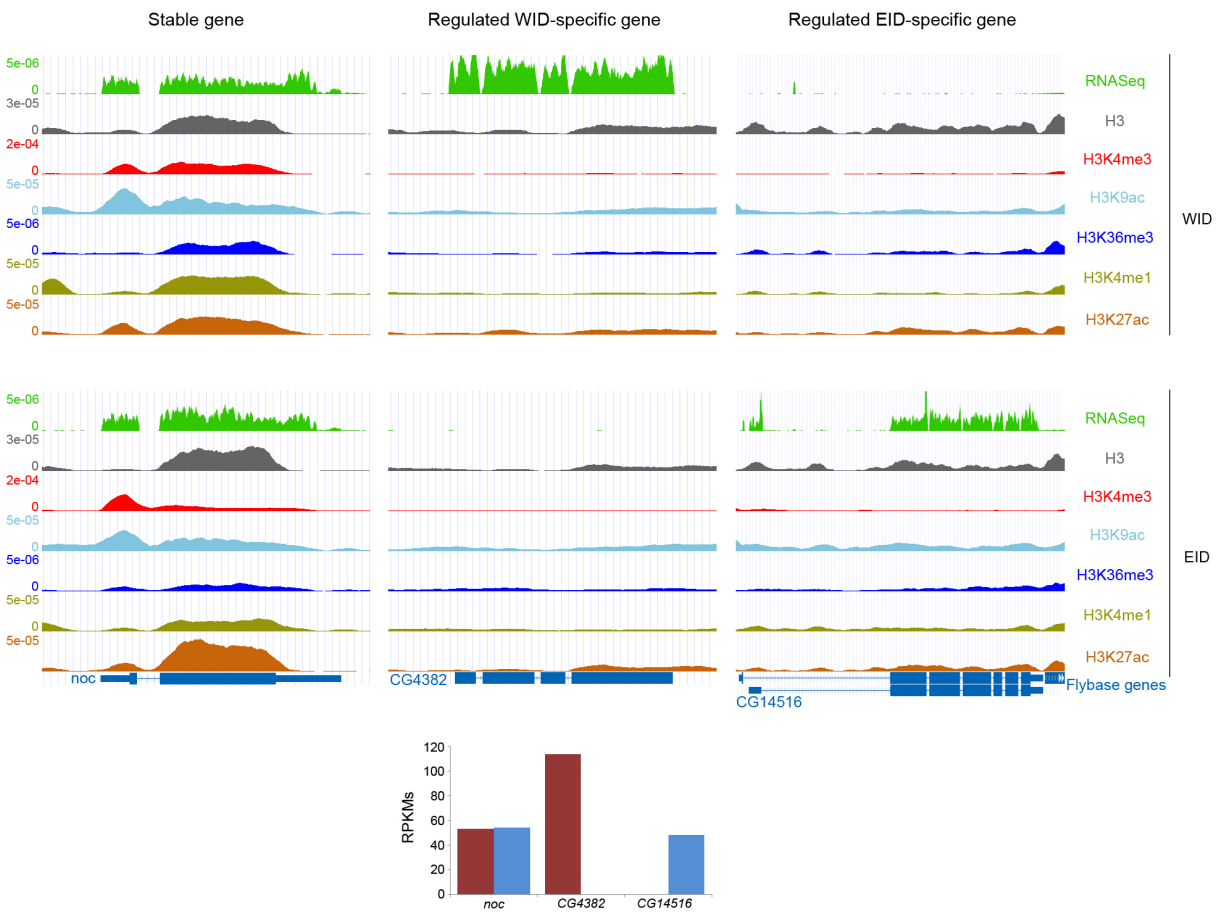

Stable gene
Regulated WID-specific gene
Regulated EID-specific gene
RNASeq
H3
H3K4me3
H3K9ac
H3K36me3
H3K4me1
H3K27ac
WID
EID
noc
CG4382
CG14516
Flybase genes
RPKMs
120
100
80
60
40
20
0
noc
CG4382
CG14516

Figure 5

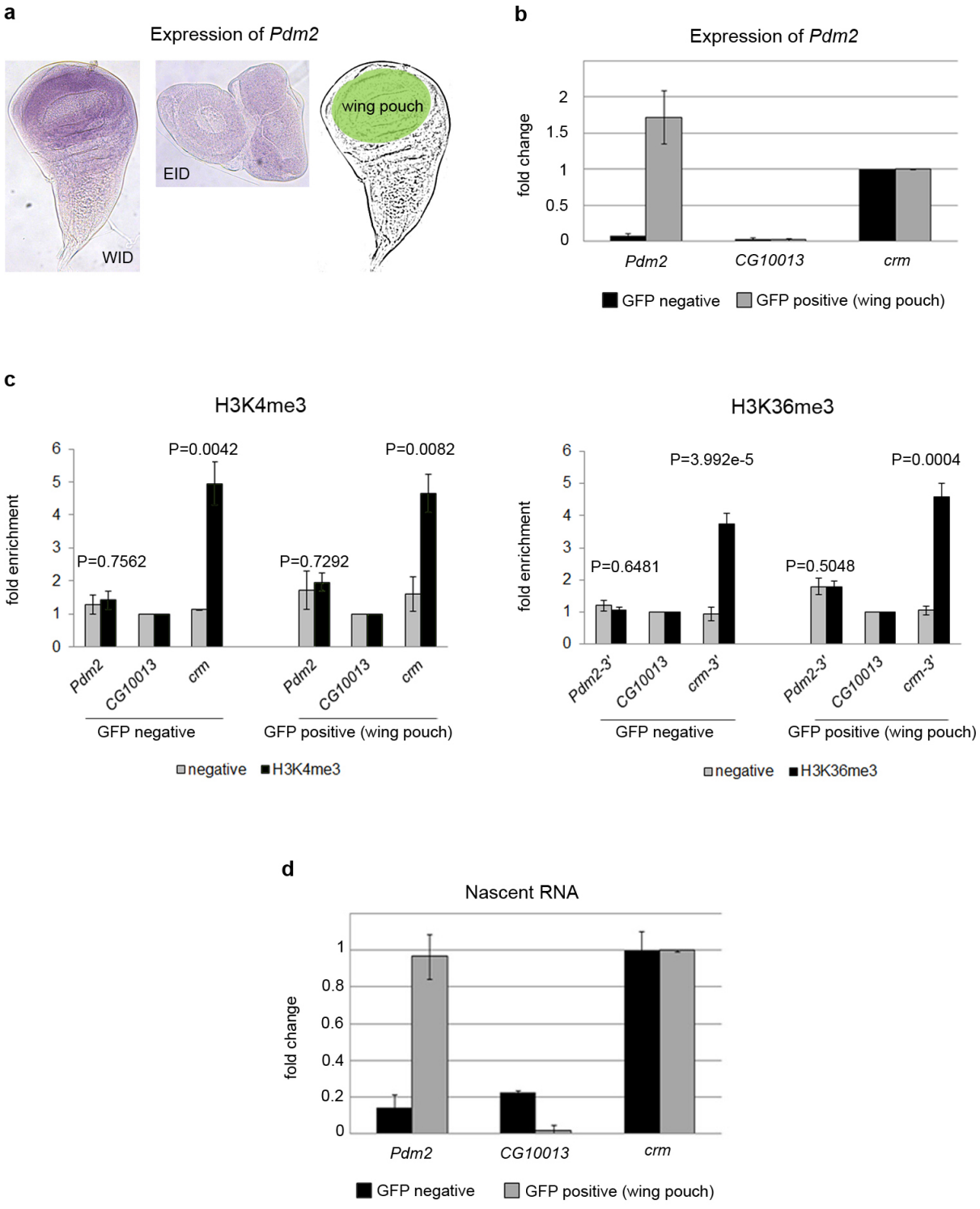

a
Expression of Pdm2
WID
EID
wing pouch
b
Expression of Pdm2
fold change
2
1.5
1
0.5
0
Pdm2
CG10013
crm
GFP negative
GFP positive (wing pouch)
c
H3K4me3
fold enrichment
P=0.7562
P=0.0042
P=0.7292
P=0.0082
Pdm2
CG10013
crm
GFP negative
GFP positive (wing pouch)
negative
H3K4me3
H3K36me3
P=0.6481
P=3.992e-5
P=0.5048
P=0.0004
Pdm2-3'
CG10013
crm-3'
H3K36me3
d
Nascent RNA
fold change
1
0.8
0.6
0.4
0.2
0
Pdm2
CG10013
crm
GFP negative
GFP positive (wing pouch)

Figure 6

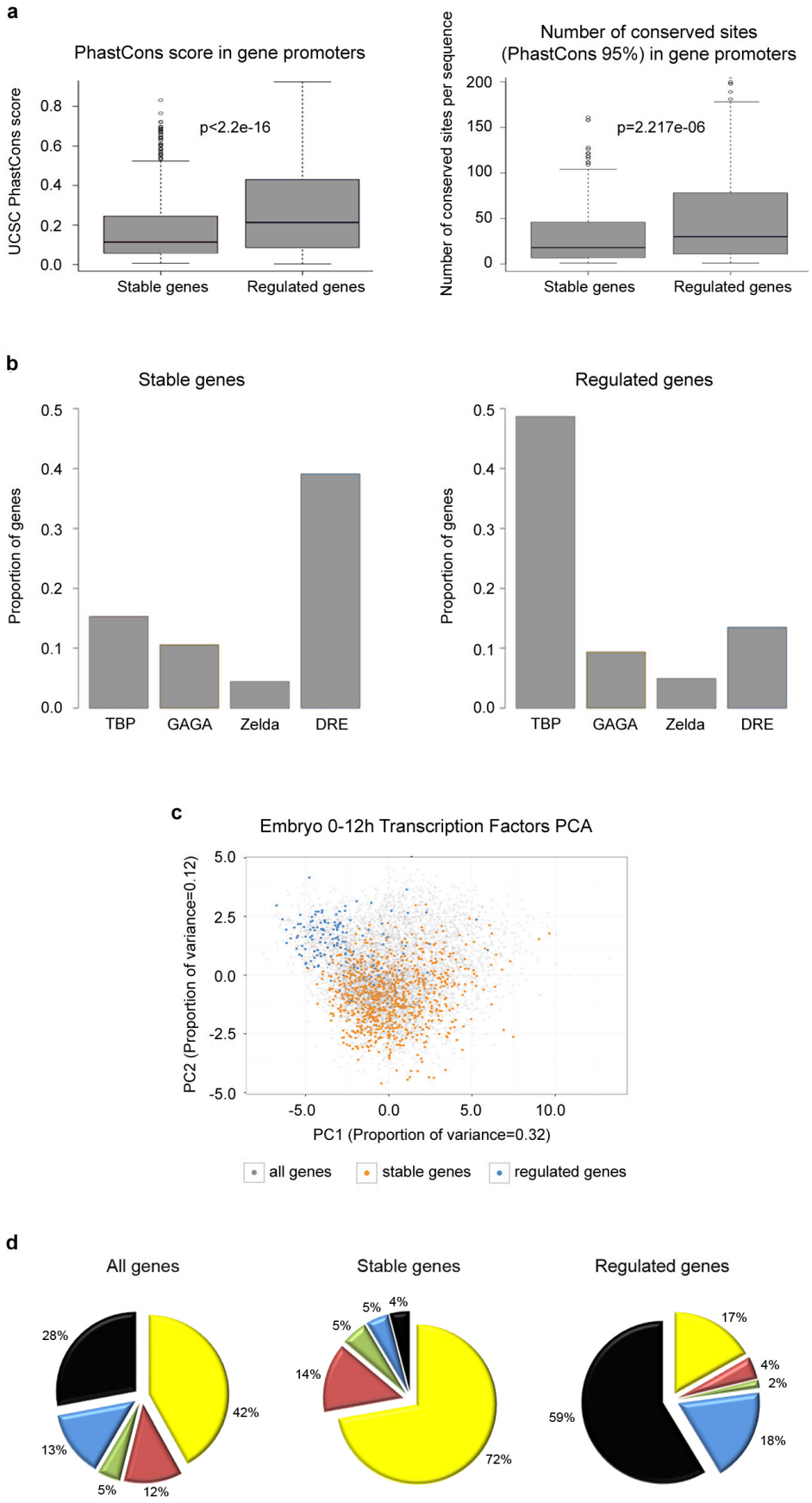

a
PhastCons score in gene promoters
UCSC PhastCons score
p<2.2e-16
Stable genes
Regulated genes
Number of conserved sites
(PhastCons 95%) in gene promoters
Number of conserved sites per sequence
p=2.217e-06
Stable genes
Regulated genes
b
Stable genes
Proportion of genes
TBP
GAGA
Zelda
DRE
Regulated genes
Proportion of genes
TBP
GAGA
Zelda
DRE
c
Embryo 0-12h Transcription Factors PCA
PC2 (Proportion of variance=0.12)
PC1 (Proportion of variance=0.32)
all genes
stable genes
regulated genes
d
All genes
28%
42%
13%
5%
12%
Stable genes
5%
4%
5%
14%
72%
Regulated genes
17%
4%
2%
59%
18%

Figure 7

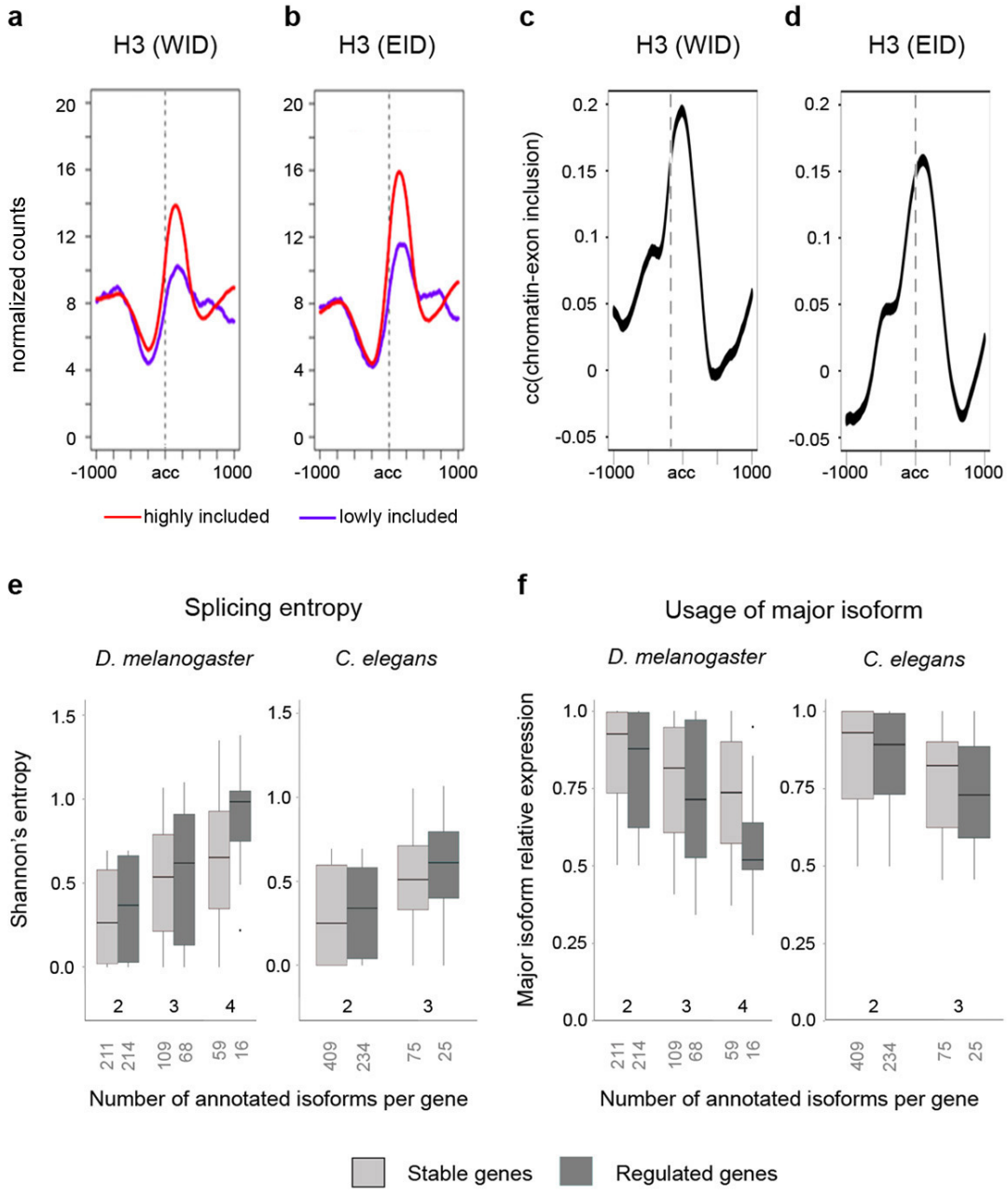

a
H3 (WID)
b
H3 (EID)
c
H3 (WID)
d
H3 (EID)
normalized counts
cc(chromatin-exon inclusion)
-1000
acc
1000
highly included
lowly included
e
Splicing entropy
f
Usage of major isoform
D. melanogaster
C. elegans
Shannon's entropy
Major isoform relative expression
Number of annotated isoforms per gene
Stable genes
Regulated genes

Figure 8

Figure captions

Figure 1: Distribution of histone modifications during fly development. **a**, Expression of stable, regulated, and silent genes during fly development at the time point of maximum expression for each gene. Gene expression was computed as FPKMs by the modENCODE consortium. **b,** Levels of H3K4me3, H3K9ac, H3K4me1 and H3K27ac at the time point of maximum expression during *D. melanogaster* development. These values represent the maximum height of the ChIPSeq peak within the body of the gene. P-values were computed using the Wilcoxon text (two-sided). **c**, Profiles of H3K4me3 during the twelve fly developmental time points in *CG8636*, a gene stably expressed during fly development, and *CG16733*, a pupa-specific gene. The expression (measured as FPKMs) along these points for these two genes is given on the left. **d**, Levels of H3K27me3 and H3K9me3 at the time point of maximum expression, computed as the average height of the ChIPSeq signal within the body of the gene, in stable, regulated and silent genes.

Figure 2: Distribution of histone modifications during worm development. a, Expression of stable, regulated and silent genes during worm development at the time point of maximum expression for each gene. **b,** Levels of H3K4me3 and H3K36me3 at the time point of maximum expression during worm development. These values represent the maximum height of the ChIP-chip peak within the body of the gene. P-values were computed using Wilcoxon text (two-sided).

Figure 3: Gene expression and histone modifications in regulated broadly expressed and stable tissue-specific genes at third instar larvae. a, Diagrams of developmentally regulated genes broadly expressed across multiple tissues at third instar-larvae L3 (left panel), and stable genes expressed in only one tissue at L3 (right panel). **b,** Gene expression levels at L3 measured by whole organism RNASeq (left panel). The number of genes in each category is given under the boxplots. Validation by qPCR of the expression at L3 of regulated broadly expressed genes compared to

a stable gene (*Bmcp*) and a silent gene (*CG5367*) (right panel). Error bars represent the Standard Error of the Mean (SEM) from three independent replicates. **c,** Levels of H3K4me3, H3K9ac, H3K4me1 and H3K27ac on whole L3 individuals. The seven regulated genes broadly expressed at L3 are depicted as red dots within the boxplots. P-values were computed using the Wilcoxon test (two-sided). **d,** Validation by individual ChIPs and qPCR of H3K4me3 and H3K9ac in regulated genes broadly expressed at L3. H3K4me3 and H3K9ac ChIPs are represented as enrichment of the marks over the silent gene (*CG5367*). Error bars represent the SEM from three independent replicates.

Figure 4: Profiles of H3 and histone modifications in Wing (WID) and Eye-antenna (EID) imaginal discs. a, Profiles on stable and silent genes in WID and in EID. **b**, Profiles on regulated WID-specific genes in WID and EID. **c,** Profiles on regulated EID-specific genes in WID and EID.

Figure 5: Profiles of RNA expression, H3 and histone modifications in Wing (WID) and Eye-antenna (EID) imaginal discs. *Noc* is a gene stably expressed in WID and EID; *CG4382* a WID-specific and *CG14516*, an EID-specific gene. Levels of gene expression (as RPKMs) are depicted at the bottom of the panels. Screenshots have been obtained through the UCSC Genome Browser[38].

Figure 6: Active transcription of *Pdm2* without chromatin modifications. a, Expression of *Pdm2* in WID (left panel) and EID (middle panel) labeled with a *Pdm2*-specific probe. The gene is only expressed in the wing pouch of the WID, highlighted in green. **b,** Expression of *Pdm2* in sorted cells analyzed by qPCR. Error bars represent the SEM from three biological replicates. **c**, ChIP analysis of H3K4me3, H3K36me3 and of negative controls without antibody on sorted cells. ChIPs are represented as enrichment of the marks over a silent gene non-marked with H3K4me3 and H3K36me3 (*CG10013*). Error bars represent the SEM from at least three biological replicates. P-values were computed using the t-test (two-

sided). **d**, Newly transcribed RNA of GFP-sorted cells. Nascent RNA is normalized by the control gene *crm*. Error bars represent the SEM of four biological replicates.

Figure 7: Promoter architecture in stable and developmentally regulated genes. a, Conservation of core promoter sequence**.** Upper panel: distribution of PhastCons scores derived from 12 *Drosophila* species in the promoter sequence (defined as 200 bp upstream of the TSS) of stable and regulated genes. Lower panel: conservation of transcription factor binding motifs. We identified the predicted binding motifs for Transcription Factors that have a PhastCons score greater than 0.95 in the promoter sequence of stable and variable genes. Boxplots show the distribution of the number of conserved motifs only for promoters that contain at least one prediction. P values were computed using Wilcoxon test (one-sided). **b,** Percentage of the promoters in stable and regulated genes that contain binding sites for TBP, GAGA, Zelda and DRE. **c,** Principal component analysis (PCA) of genes expressed in the *Drosophila* embryo between 0 and 12h, based on the ChIP-chip binding profiles at their promoters of 20 Transcription factors. **d,** Stable and regulated genes in chromatin domains segmented according to Filion et al.[27]. The BLACK corresponds mostly to repressive chromatin. YELLOW and RED chromatin contain proteins and histone modifications typical of transcriptionally active regions. GREEN and BLUE chromatin correspond to two types of repressive chromatin: heterochromatin including heterochromatin protein 1 (HP1) (GREEN) and Polycomb group (PcG)-associated chromatin (BLUE).

Figure 8: Chromatin structure and splicing. **a-b**, H3 on highly (red) and lowly (blue) included exons in WID (**a**) and EID (**b**). **c-d**, Correlation between exon inclusion and H3 across exon acceptor sites in WID (**c**) and EID (**d**). **e**, Distribution of Shannon's Entropy in stable and regulated genes. Shannon's entropy is computed at the time point during development in which gene expression is the maximum. The number of genes of each category appears below the X-axis in grey. **f**, Distribution of the relative

usage of the major isoform for genes with different number of isoforms. The Y-axis is the fraction of the total transcriptional output of the gene that is captured by the most abundant isoform.

Results – Part II

Research

Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs

Hagen Tilgner,[1,3] David G. Knowles,[1] Rory Johnson,[1] Carrie A. Davis,[2] Sudipto Chakrabortty,[2] Sarah Djebali,[1] João Curado,[1] Michael Snyder,[3] Thomas R. Gingeras,[2] and Roderic Guigó[1,4]

[1]*Centre for Genomic Regulation (CRG) and UPF, E-08003, Barcelona, Catalonia, Spain;* [2]*Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA;* [3]*Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA*

Splicing remains an incompletely understood process. Recent findings suggest that chromatin structure participates in its regulation. Here, we analyze the RNA from subcellular fractions obtained through RNA-seq in the cell line K562. We show that in the human genome, splicing occurs predominantly during transcription. We introduce the coSI measure, based on RNA-seq reads mapping to exon junctions and borders, to assess the degree of splicing completion around internal exons. We show that, as expected, splicing is almost fully completed in cytosolic polyA+ RNA. In chromatin-associated RNA (which includes the RNA that is being transcribed), for 5.6% of exons, the removal of the surrounding introns is fully completed, compared with 0.3% of exons for which no intron-removal has occurred. The remaining exons exist as a mixture of spliced and fewer unspliced molecules, with a median coSI of 0.75. Thus, most RNAs undergo splicing while being transcribed: "co-transcriptional splicing." Consistent with co-transcriptional spliceosome assembly and splicing, we have found significant enrichment of spliceosomal snRNAs in chromatin-associated RNA compared with other cellular RNA fractions and other nonspliceosomal snRNAs. CoSI scores decrease along the gene, pointing to a "first transcribed, first spliced" rule, yet more downstream exons carry other characteristics, favoring rapid, co-transcriptional intron removal. Exons with low coSI values, that is, in the process of being spliced, are enriched with chromatin marks, consistent with a role for chromatin in splicing during transcription. For alternative exons and long noncoding RNAs, splicing tends to occur later, and the latter might remain unspliced in some cases.

[Supplemental material is available for this article.]

Central in the pathway leading from primary transcripts to mature functional RNAs is splicing, the process by which intervening sequences in the primary transcript (introns) are excised and the remaining sequences (exons) are concatenated together to form the mature eukaryotic RNAs. Conserved sequence motifs, the splice sites, mark exon–intron boundaries and are recognized by elements of the splicing machinery. Splice site sequences, however, do not carry enough information to unequivocally specify exon–intron boundaries, and a plethora of other sequence motifs, recognized by a variety of RNA binding proteins, contribute to define and regulate splice site selection (Graveley 2000; Smith and Valcárcel 2000; Wang and Burge 2008). While there have been considerable advances in modeling splicing from features in the primary transcript sequence (Wang et al. 2004; Barash et al. 2010), it is currently close to impossible to predict from the analysis of mammalian primary RNA sequence alone neither the entire exon–intron structure of transcripts nor their tissue specific expression pattern (i.e., the abundance of given transcript in a given cell type).

It appears thus that other factors, not necessarily encoded in the sequence of the primary transcript, may play a role in splicing definition. Indeed, there is a growing body of evidence suggesting that chromatin structure could play a role in splicing. A number of reports have demonstrated that eukaryotic exonic sequences are enriched in positioned nucleosomes and that some histone modifications show characteristic exonic patterns (Andersson et al. 2009; Hon et al. 2009; Kolasinska-Zwierz et al. 2009; Nahkuri et al. 2009; Schwartz et al. 2009; Spies et al. 2009; Tilgner et al. 2009). Intragenic histone modifications and chromatin structure influences on alternative splicing events have been documented in detail for the fibronectin and *FGFR2* gene (Allo et al. 2009; Schor et al. 2009; Luco et al. 2010), and CTCF-mediated local RNA polymerase II (Pol II) pausing has been shown to influence alternative splicing (Shukla et al. 2011). While the underlying molecular mechanisms connecting chromatin structure with splicing are largely unknown, they require, in principle, for splicing to be somehow connected to transcription. That splicing can be carried out during transcription has been known for a long time (Beyer and Osheim 1988), and increasing evidence exists of coupling between transcription and splicing (Cramer et al. 1997; Roberts et al. 1998; Kadener et al. 2001; Nogues et al. 2002; de la Mata et al. 2003; Howe et al. 2003). Intron removal during transcription has been shown to be predominant in the intron-poor genome of *Saccharomyces cerevisiae* (Carrillo Oesterreich et al. 2010), and recently, Ameur et al. (2011) have proposed co-transcriptional splicing to be widespread in the human brain, based on the analysis of whole-cell, total RNA sequencing (RNA-seq). Here, we analyze the RNA that is still residing on the chromatin template,

[4]Corresponding author
E-mail roderic.guigo@crg.cat
Article and supplemental material are at http://www.genome.org/cgi/doi/10.1101/gr.134445.111. Freely available online through the *Genome Research* Open Access option.

22:1616–1625 © 2012, Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/12; www.genome.org

as well as nuclear and cytosolic RNA in its polyadenylated and nonpolyadenylated form. This fractional approach, including separate RNA-seq in the cytosol, enables us to define splicing completion at multiple nuclear stages. Thus we can define the fraction of splicing events around an exon that is co-transcriptional. This, in turn allows the definition of smaller subsets of candidates "with a tendency for post-transcriptional splicing"(postTS) without being biased by intron-retention events. We show that co-transcriptional splicing is predominant in the human genome, providing the basis for the understanding of the role of chromatin structure in splicing definition and regulation. We investigate the rules that determine whether the introns around an exon are to be spliced co-transcriptionally. We also observe a 5′-to-3′ trend in splicing completion, which causes more downstream splicing events to be more prone to postTS and makes the distance to the polyA-site one of the most important factors in determining when splicing is completed. This 5′-to-3′ trend is countered by shorter introns and stronger splice site strengths toward the end of the gene—two features that we have found to promote early, co-transcriptional splicing. Further significant predictors of co-transcriptional splicing include the rate of gene transcription and covalent histone modifications. Long noncoding RNAs (lncRNAs) appear to be less efficiently spliced than protein coding genes and, on occasion, may even remain unspliced. Exons, for which the surrounding introns are in the process of being spliced, are enriched with chromatin marks, consistent with a role for chromatin in splicing during transcription, and splicing around alternative exons is, on average, more post-transcriptional than for constitutive exons.

Results

Deep RNA-seq of subcellular fractions

We have used deep RNA-seq, performed within the framework of the ENCODE project, to interrogate with unprecedented resolution distinct RNA fractions from a number of cellular compartments in human immortalized myelogenous leukemia cells K562: chromatin-associated total RNA, polyA− and polyA+ nuclear RNA, as well as polyA− and polyA+ cytosolic RNA (Fig. 1A; see Djebali et al. 2012 and Supplemental Information/Methods for details on RNA fractionating and controls, sequencing, and bioinformatic analysis of the sequence data). Monitoring of these compartments provides snapshots of the different stages of RNA processing within the cell.

The completed splicing index

We introduce a measure, based on the RNA-seq reads mapping to the exon junctions, to assess the degree of completion of splicing around internal exons. We simply count the number of reads mapping across the exon boundaries into the adjacent intron sequence (which originate from primary, unspliced mRNA molecules), as well as the number of reads split-mapping across exon–exon junctions, either from the exon to another exon of the same gene or between an upstream and a downstream exon (both types of read originating from a successfully completed splicing event) (Fig. 1B). Based on these numbers, we compute the completed splicing index (coSI) of a given exon, corresponding thereby to the weighted percentage of reads supporting splicing completion

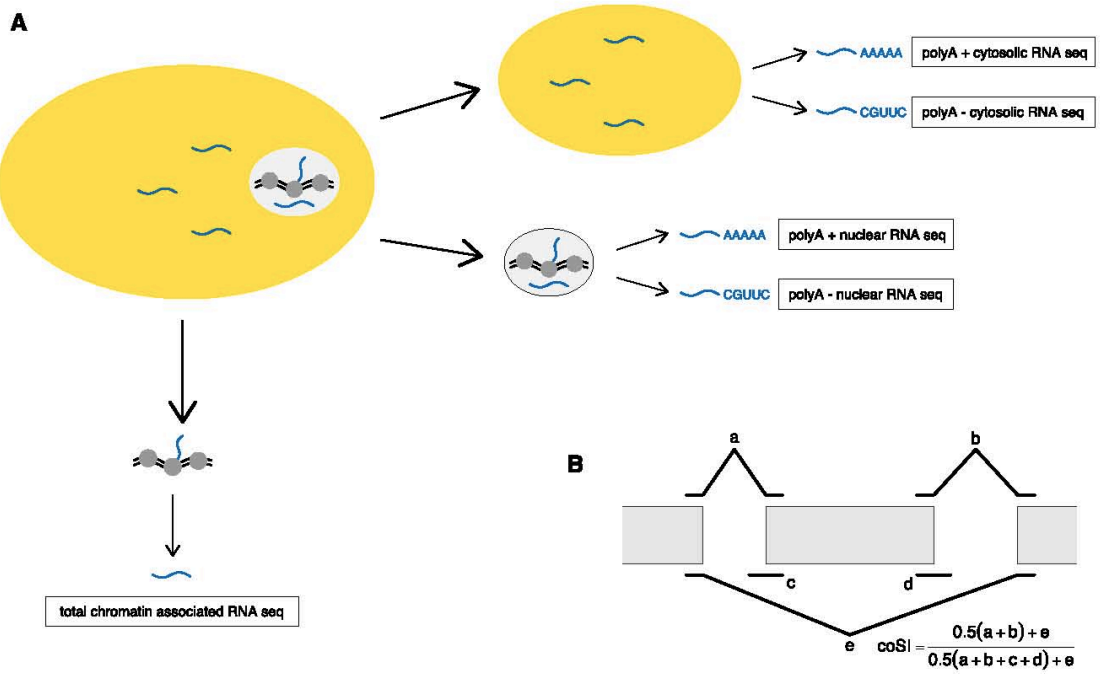


Figure 1. (*A*) Long RNA-seq data sets used in this analysis. (*B*) Definition of the completed splicing index (coSI) for each internal exon and each RNA-seq data set.

around the exon. The coSI value can be broadly assumed to correspond to the fraction of exon-containing RNA molecules in which splicing in the region around the exon has already been carried out. A coSI value of 1 means entirely completed splicing, while coSI = 0 indicates that the exon is still completely included in the sequence of the primary transcript.

Most human exons are already partially spliced in chromatin-associated RNA

We have computed coSI scores for human internal exons (see Supplemental Methods) in all analyzed K562 RNA fractions (Supplemental Table S1). We have observed a higher correlation of coSI values (R = 0.82) between two replicates of the chromatin fraction than between the chromatin fraction and other fractions (R of between 0.35 and 0.71) (Supplemental Fig. S1A–G), confirming that overall coSI values are reproducible within a given experiment.

Figure 2 shows the distribution of coSI scores in the different RNA fractions that we have interrogated. As expected, for most exons, splicing of the corresponding introns is fully completed in the cytosolic polyA+ fraction (92% of the exons have a coSI ≥ 0.95), as well as the cytosolic polyA− fraction (data not shown). Of even more interest with respect to splicing is the polyA− nuclear fraction, in which the median coSI is 0.84. For 16% of the exons, their surrounding introns are completely spliced in this fraction, and only for a vanishing fraction (<0.2% with coSI ≤ 0.05) do the corresponding introns remain completely unspliced. The polyA− nuclear fraction contains RNA molecules of three types: first, RNAs that are still being transcribed and for which transcription has not yet reached the polyA-site; second, RNAs that have been released from the transcribing Pol II, before it could reach the polyA-site; and, third, products of aborted transcription. The high degree of splicing completion in this fraction therefore suggests that splicing is mostly initiated before completion of transcription. Even more enriched for RNAs in the act of being transcribed is the chromatin-associated fraction. With a median coSI of 0.75 in this fraction, around most exons we see large amounts of completed splicing. For 5.6% of the exons, we see absolutely completely spliced introns (coSI ≥ 0.95); however, as in the polyA− nuclear fraction, only a tiny fraction of exons (<0.3%, coSI ≤ 0.05) are surrounded

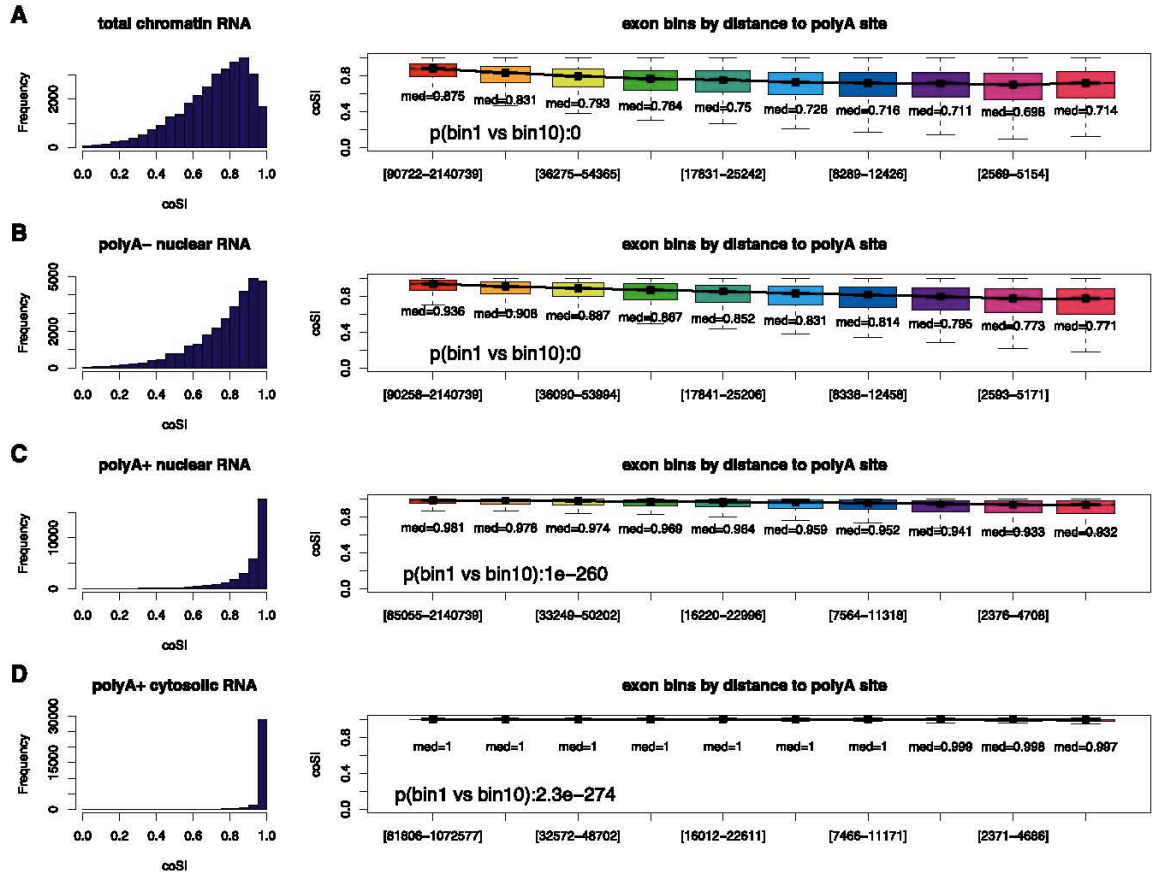


Figure 2. Histogram of coSI values (*left*) and boxplots of coSI values in bins according to the distance of an exon to the annotated polyA site (intervals on *x*-axis give minimum and maximum distance in each bin; *right*) for the total chromatin-associated RNA fraction (*A*), the polyA− nuclear fraction (*B*), the polyA+ nuclear fraction (*C*), and the polyA+ cytosolic fraction (*D*). *P*-values were calculated comparing the first and the last bin, using a two-sided Wilcoxon rank sum test. Numbers *below* boxplots indicate the median value of the according distribution.

by completely unspliced introns, further suggesting that splicing is intimately coupled and occurs almost simultaneously with transcription.

In order to exclude the possibility that large numbers of reads mapping to junctions are in fact artifactual, we have shifted the annotation by 30 bp against the transcription direction, so that in this "fake annotation," no junction is real, although practically all exons still lie within genes. As a result, the number of reads mapping to junctions has dropped 29-fold (from 11.4 million to less than 400,000), and these fake-junction reads are highly enriched in reads with two mismatches (62% compared with 7.5% for the real annotation), suggesting that these fake-junction reads contain larger numbers of false mappings. We have therefore compared chromatin-fraction coSI values for the real and for the fake annotation, using only reads (whether mapping to junctions or genome) without mismatches. The median coSI for the real annotation is 0.71, while the median coSI for the fake annotation is 0.00 (Supplemental Fig. S1H), proving that our results are not flawed by false junction mappings.

Consistent with a general strong coupling of splicing and transcription, we have found that coSI values decrease with decreasing distance to the polyA site, pointing to a "first transcribed, first spliced" trend (Fig. 2). The trend is very strong for the chromatin-associated and polyA− nuclear RNAs and is weak or absent for the polyA+ RNAs, and we have made similar observations when using an acceptor-based-definition of completed splicing (Supplemental Fig. S2). We have observed the same trend when combining all genes together and normalizing coSI scores of exons to their relative position within the gene (Supplemental Fig. S3). In these idealized genes, the coSI reaches its maximum between 20% and 30% of the gene length, indicating that splicing of introns very near the 5′ end of the gene could require a little more time. Yet from ~20%–30% of the transcript, the coSI decreases gradually. This decrease is less strong than in Figure 2, supposedly because long and short genes are considered equally, thereby minimizing the influence of exons that are very far from the polyA site.

We have specifically examined the coSI scores of exons from two genes that constitute well-studied examples of co-transcriptional splicing—the fibronectin (Cramer et al. 1997; Kadener et al. 2001; Nogues et al. 2002; de la Mata et al. 2003; Pandya-Jones and Black 2009) and *SRC* (Pandya-Jones and Black 2009) genes, and we have found that their exons have higher coSI values in the chromatin and nuclear polyA+ fractions compared with exons of other genes (Supplemental Fig. S4). This finding is not a pure consequence of gene length, as exons of 4000 longer genes show lower coSI values than fibronectin and *SRC* exons (Supplemental Fig. S4). The above observations (see Fig. 2; Supplemental Fig. S4) are not caused by incomplete annotation. Indeed, when split-mapping unmapped reads from the total chromatin fraction and excluding all exons that are within 250 bp of a potential novel splice site, we observe essentially the same trend (Supplemental Fig. S5).

Analysis of reads mapping to the genome (including exonic reads and reads deep within the introns, both of which were not used for the coSI calculation) confirms that high coSI values correspond to exons whose surrounding introns are mostly removed. Indeed, exons with low coSI in chromatin RNA show almost flat RNA-seq profiles in this RNA fraction, whereas high coSI exons show strong RNA-seq peaks on exons (Supplemental Methods; Fig. 3A,B). Exons with very low coSI values in the chromatin fraction seem to correspond to exons whose surrounding introns are spliced later, even after polyadenylation, whereas introns surrounding exons with medium or high coSI values in the total chromatin fraction seem to be spliced early, as the former show intronic reads, whereas the latter do not, in the polyA+ nuclear fraction (Fig. 3C). As expected, all exon groups show almost no evidence for intronic unspliced reads in the cytosolic polyA+ fraction (Fig. 3D). When these profiles are normalized for cytosolic polyA+ gene expression, the peak height of all three exon bins is essentially identical in the cytosolic polyA+ fraction (Supplemental Fig. S6), indicating that this observed peak height is characteristic for completed splicing. Consistent with the 5′-to-3′ coSI bias, we have found higher intronic compared with exonic read-depth as exons are closer to the polyA site (Supplemental Fig. S7).

To further rule out the possibility that our observations may originate from technical artifacts, we have analyzed CAGE tags and antisense reads, in addition to clustering the subnuclear fractions according to coSI scores. The results strongly argue against our observations originating from technical artifacts (Supplemental Information; Supplemental Figs. S8, S9).

Gene coSI values

In order to complement the exon-based view of splicing completion, we have computed coSI values at the gene level, as a function of the number of reads mapping to intron–exon junctions within the gene and of the number of reads split mapping between exons from the gene (Supplemental Information and Methods). We have found the median gene coSI value in the total chromatin fraction to be 0.618, again supporting the idea that a majority of splicing events is carried out co-transcriptionally.

Spliceosomal RNAs are enriched in chromatin-associated RNA

If splicing occurs mostly co-transcriptionally and therefore in proximity to the chromatin template, one would expect that RNAs of the splicing machinery would also reside in proximity to chromatin. We have therefore investigated the subcellular location of U1-U6 and U6atac (UxRNAs) based on RNA-seq of small RNAs performed in five different subcellular locations (Nucleus, Cytosol, Nucleoplasm, Nucleoli, and Chromatin) (Djebali et al. 2012; Supplemental Methods). As predicted, all spliceosomal UxRNAs—that is, U1, U2, U4, U5, U6, and U6atac, but not U3—are clearly enriched in the chromatin-associated fraction compared with the other fractions (Fig. 4A,B,D–G). In contrast, U3 and snoRNAs (excluding U RNAs), both of which are thought not to be involved in splicing, were highly enriched in the nucleoli fraction (Fig. 4C,H), as expected from their known functions. Of special interest in this respect is the observation that U6atac, a spliceosomal RNA of the minor spliceosome, is also enriched in the chromatin fraction. This strongly suggests two things: First, the minor spliceosome, similar to the major spliceosome, is assembled co-transcriptionally; and, second, if co-transcriptional spliceosome assembly is an attribute of both spliceosome systems, it appears likely that both types of intron removal occur in the same way, namely, co-transcriptionally in most cases.

Exon coSI values correlate with features of gene, exon, and chromatin structure

A number of sequence features characterizing the exons and their surrounding regions seem to weakly correlate with exon coSI values in chromatin (Supplemental Fig. S10). The most notable correlation is with distance to the PolyA site (see also Fig. 2) and, albeit

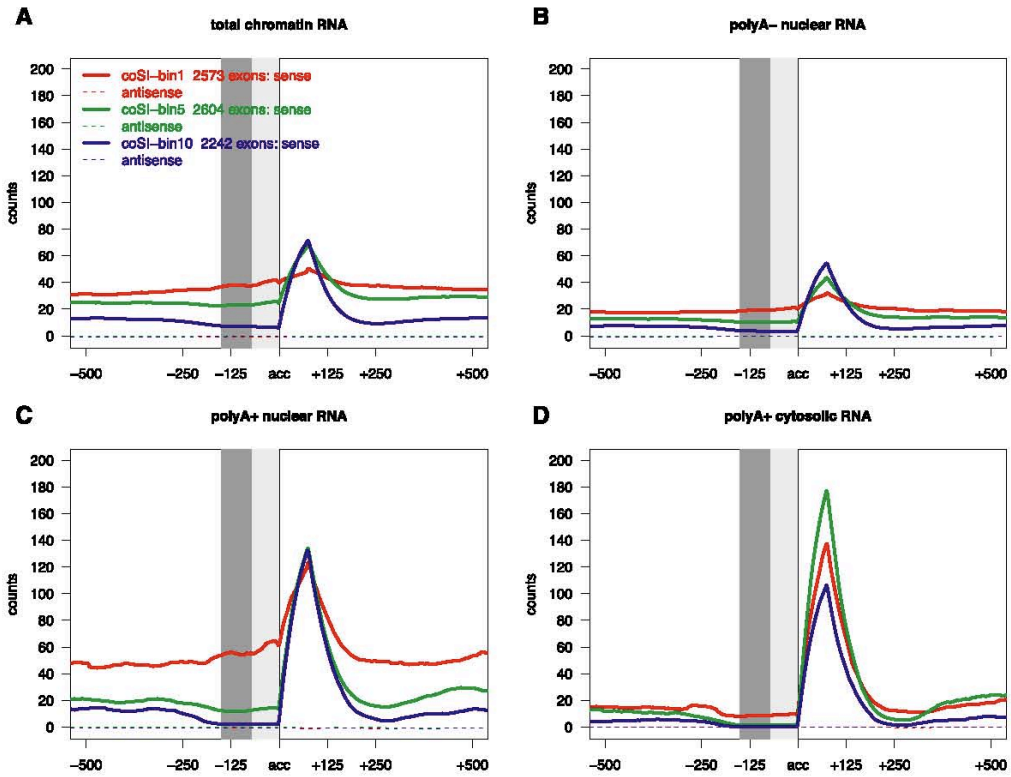


Figure 3. RNA-seq profile plots using RNA-seq reads mapping to the genome only, aligned at the acceptor. At each aligned position, the average number of overlapping RNA-seq reads (mapping to the genome) for all exons in each bin (according to coSI values in the total chromatin-associated RNA fraction in all four subfigures) is plotted for sense (solid lines) and antisense strand (dashed lines). Here, only exons that are at least 150 bp away from any other exon are used. RNA-seq profiles for the total chromatin-associated RNA fraction (*A*), the polyA– nuclear fraction (*B*), the polyA+ nuclear fraction (*C*), and the polyA+ cytosolic fraction (*D*). (Dark gray area) Positions that are guaranteed to be covered only by reads that were not used for the coSI value calculation; (both gray areas) positions that are guaranteed to be intronic. Note that these profiles are not normalized for gene expression. We added profiles normalized for cytosolic polyA+ gene expression in Supplemental Figure S6.

somehow weaker, with the distance to the transcription start site (TSS). In addition, exon coSI values correlate positively with the strength of the acceptor sites and GC content and anti-correlate with the length of the downstream intron—this is supposedly because reads spanning the exon–intron border can be observed once the donor is transcribed, while splicing can only be carried out once the entire downstream intron is transcribed. It also appears that the presence of binding sites for some splicing factors weakly correlate with coSI scores (data not shown). We have further investigated the exonic behavior of a number of chromatin modifications (Ernst et al. 2011, monitored through ChIP-seq in K562, see Supplemental Information and Methods) depending on the exon coSI value in chromatin-associated RNA. All chromatin marks monitored, as well as nucleosome (Kundaje et al. 2012) and Pol II occupancy, negatively correlate with chromatin coSI values (Supplemental Fig. S10). That is, there is a general enrichment of chromatin marks in exons with low coSI values, consistent with the DNA in these exons being still in chromatin status before or during transcription.

In order to understand how all these factors may contribute to co-transcriptional splicing, we have built a linear model in which exon coSI values in the chromatin are predicted from these factors. The linear model using all 84 variables (Supplemental Methods) achieves a correlation coefficient (cc) of 0.48 (Fig. 5A), comparing observed and predicted coSI values. The distance of the exons from annotated TSSs and polyA-sites are the most informative variables (cc of ~0.31) (Fig 5B). Acceptor and donor strength, as well as length of the surrounding introns and the exon itself, GC content of the exon, gene expression in the nuclear polyA+ fraction, chromatin status and marks, and binding sites for splicing factors progressively add more information to the prediction of coSI values (Fig. 5B–E).

coSI values reflect contrasting patterns of splicing dynamics

Analysis of coSI values across fractions reveals the specific processing pattern of the RNA in the vicinity of exons (Fig. 6A). Indeed, for 94% of the exons, the coSI value shows a monotonically increasing behavior from the total chromatin through the polyA+ nuclear to the polyA+ cytosolic fraction. Clustering according to coSI values of exons of the RNA fractions indicates that there is a large population of exons that are rapidly spliced in the chromatin fraction, while, at the other end, there is a smaller population of exons that appear to delay completion of splicing even after polyadenylation, just before exporting to the cytosol (Fig. 6A). We have specifically investigated the characteristic traits of exons that appear to delay splicing post-transcriptionally. Thus, we have arbitrarily selected

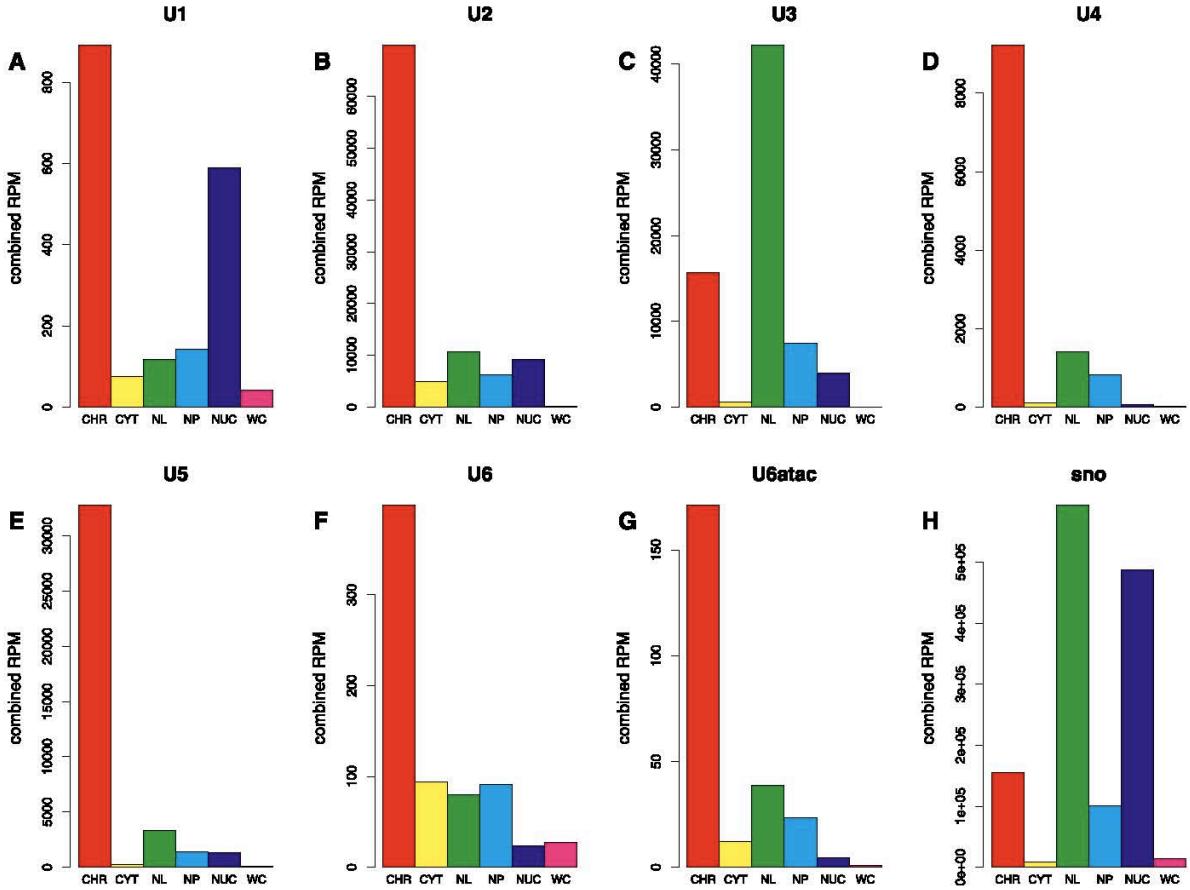


Figure 4. An RPM was calculated based on short RNA-seq in each subcellular fraction—total chromatin fraction (CHR; red), total cytoplasmic fraction (CYT; yellow), total nucleoli fraction (NL; green), total nucleoplasmic fraction (NP; light blue), total nuclear fraction (NUC; purple), total whole-cell fraction (WC; pink)—and summed for all genes encoding for U1-RNA (*A*), U2-RNA (*B*), U3-RNA (*C*), U4-RNA (*D*), U5-RNA (*E*), U6 RNA (*F*), U6atac (*G*), and non–U-RNA snoRNAs (*H*).

1390 exon candidates "with a tendency for postTS (postTS-exons)," as those with low coSI values (≤0.75) in the polyA+ nuclear fraction but high coSI values in the cytosolic polyA+ fraction (≥0.95) (Supplemental Methods). We have found that this set of postTS-exons contains 1.5-fold more alternative exons than expected by chance ($P < 3.5 \times 10^{-7}$) (Fig. 6C; see Supplemental Information and Supplemental Figures S13 and S14 for details on how alternatively spliced exons were selected). The set of postTS-exons, on the other hand, is slightly but significantly depleted of protein coding exons ($P < 5.6 \times 10^{-15}$) (Fig. 6B) and, consequently, is enriched in UTR exons—in particular in 5′UTR exons (41 out of 480 exons that are entirely within the 5′UTR, two-sided fisher *P*: 1.3×10^{-4}). Since this observation is in apparent contradiction with the general trend of lower coSI values toward the 3′end of the gene and also because the maximum coSI was not reached at the very beginning of the gene (Fig. S3), we have specifically investigated coSI values as a function of exon order. We have found that decreasing coSI values with increased distance (and exon order) from the TSS is only valid from the third exon on (Supplemental Fig. S11), with the second exon having slightly lower coSI values than the third exon, suggesting that the first intron is removed more slowly. This lower coSI value is paralleled by a trend for very long first introns and gradually decreasing intron size along the gene (Supplemental Fig. S12). Consistent with this interpretation, it is also known that acceptor and donor strength increase with distance to the TSS (Spies et al. 2009). An influence of the 5′ CAP (O'Mullane and Eperon 1998) could also contribute to first introns being spliced differently from other introns. All of these features would result in slower intron removal of the first intron and in lower coSI values of the second exon (Supplemental Fig. S11). On the 3′end of the gene, on the other hand, it appears that co-transcriptional intron removal, although disfavored by proximity to the polyA-site, is favored by shorter introns and stronger splice-sites, so that postTS remains relatively rare.

The observation that splicing dynamics differ between protein coding and noncoding exons has prompted us to specifically investigate the splicing dynamics of lncRNAs (see Supplemental Methods; Derrien et al. 2012). We find that coSI values of lncRNA exons, as a class, but also those of well-investigated lncRNAs (*H19*, *XIST*, *U50HG_SNHG5*) are dramatically lower than those of coding exons in the total chromatin fraction (Fig. 6D). Also in terms of completed splicing of the entire RNA, that is, on the gene-coSI

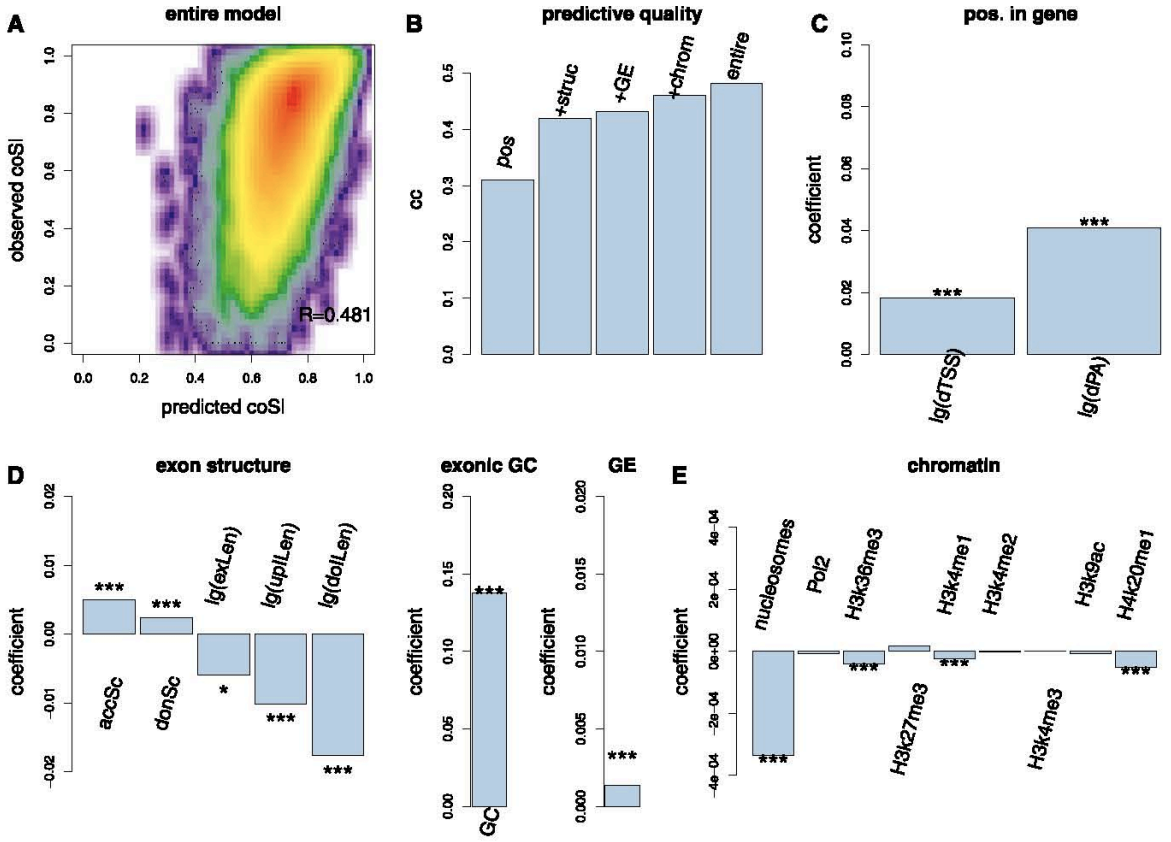


Figure 5. Linear model connecting exon-coSI values to gene, exon, and chromatin structure variables. (*A*) Smoothed scatterplot and correlation between predicted coSI values and measured coSI values using the entire model. (*B*) Correlation of predicted coSI values and measured coSI values using four increasing subsets of variables and the entire model: model with distance to TSS and distance to polyA site (pos); model additionally including acceptor strength, donor strength, log-exon-length, log-upstream-intron-length, log-downstream-intron-length and exonic GC content (+struc); model additionally including gene RPKMs from polyA+ nuclear RNA (+GE); model additionally including ChIP-seq related variables (+chrom); model including all variables (entire). (*C*) Coefficients in the entire model of distance to the TSS and to the polyA-site. (*D*) Acceptor strength (accSc), donor strength (donSc), exonic GC content (GC), log-exon-length [lg(exLen)], log-upstream-intron-length [lg(upILen)], log-downstream-intron length [lg(doILen)] and gene RPKMs from polyA+ nuclear RNA (GE). (*E*) MNase and histone modification values as described in Figure S10.

level, lncRNAs show lower splicing completion then mRNAs in the total chromatin fraction (Fig. 6E). The difference between lncRNA exons and mRNA exons persists in the nuclear polyA+ fraction (Fig. 6F), arguing that lncRNAs are often spliced later and sometimes might even not be spliced at all. This is consistent with reports that some lncRNAs remain predominantly unspliced, for example, *AIRN* and *KCNQ1OT1* (Sleutels et al. 2002; Mancini-Dinardo et al. 2006).

Discussion

Co-transcriptional splicing has recently been shown to be widespread in the intron-poor genome of *S. cerevisiae* (Carrillo Oesterreich et al. 2010). In higher eukaryotes, co-transcriptional splicing has been documented in detail for a few individual genes such as fibronectin and *SRC* (Cramer et al. 1997; Kadener et al. 2001; de la Mata et al. 2003; Pandya-Jones and Black 2009), and this mode of intron removal has been proposed to be widespread in the human brain, based on analysis of whole-cell, total RNA-seq (Ameur et al. 2011). While we coincide on the claim of widespread co-transcriptional splicing, our approach of analyzing RNA-seq data in a variety of fractions provides major advantages: First, we are able to clearly separate cytosolic RNAs, nuclear RNAs, as well as a special subset of the latter, RNAs that still reside on the chromatin template. Thus we demonstrate that spliced reads and exonic reads in the latter fraction are not the result of completely spliced RNAs, which still remain in cytosol (or nucleus) and not on the chromatin. Second, we can define, for single exons, what proportion of their surrounding introns is removed co-transcriptionally or after polyadenylation while controlling for intron retention with cytosolic RNA-seq. This we can achieve, by introducing an exon-based measure of splicing completion: This measure, the coSI, shows that most introns initiate splicing while the RNA is still associated with the chromatin—strongly suggesting that co-transcriptional splicing is also the dominant mode in the human genome. Consistent with this, we have found significant enrichment of spliceosomal snRNAs in chromatin-associated RNA compared with other cellular RNA fractions and other nonspliceosomal snRNAs. This supports the idea that exons, around which we detect a tendency for postTS, might already have been committed to splicing co-transcriptionally. Al-

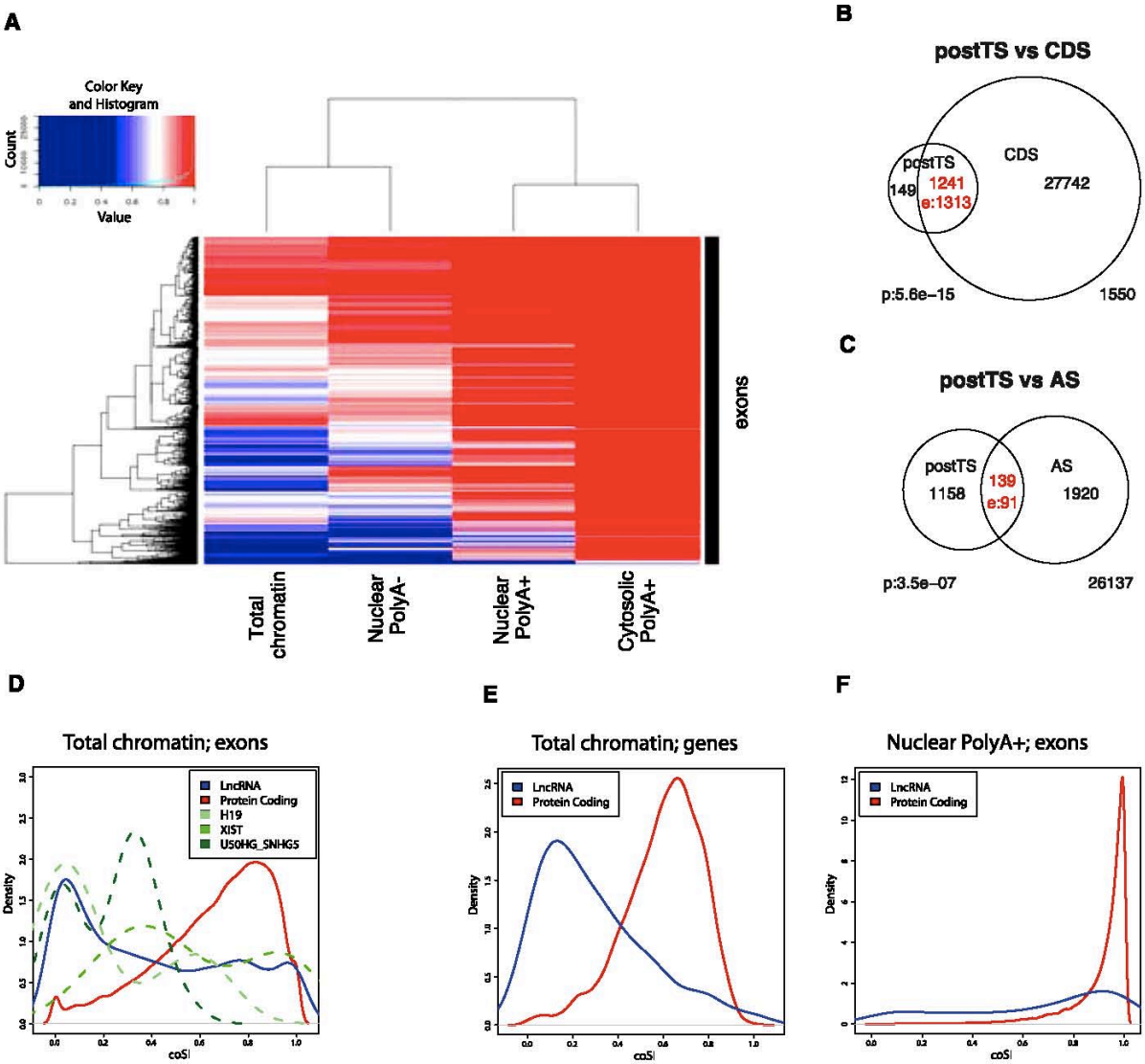


Figure 6. (*A*) Clustering of subcellular RNA fractions and exons according to exonic coSI values using four RNA fractions. From *left* to *right*: total chromatin-associated RNA, polyA− nuclear RNA, polyA+ nuclear RNA, polyA+ cytosolic RNA. Note that the scale is only linear from coSI ≥ 0.5 on. (*B*) Overlap between exons with a tendency for post-transcriptional splicing (postTS) and entirely coding exons (CDS). (*C*) Overlap between cell type specifically included AS-exons and exons with a tendency for postTS. (*D*) The distribution of coSI scores for various exon sets is shown, based on calculations for chromatin total RNA. Information is plotted for the 4933 lncRNA exons and 372,306 protein-coding gene exons that have sufficient RNA-seq reads to calculate a confident coSI score. In addition, we extracted exon values for three known lncRNAs: *H19* (18 exons), *XIST* (19 exons), *U50HG-SNHG5* (22 exons). The difference between lncRNA and protein exon coSI values is statistically significant (Wilcox test; $P < 2.2 \times 10^{-16}$). (*E*) Gene-level coSI scores from chromatin total RNA are plotted for 92 lncRNAs and 4066 protein-coding genes. The difference between the distributions is statistically significant (Wilcox test; $P < 2.2 \times 10^{-16}$). (*F*) Exon-level coSI scores from nuclear polyA+ RNA are plotted for 206 lncRNA exons and 32,496 protein-coding exons. The difference between the distributions is statistically significant (Wilcox test; $P < 2 \times 10^{-16}$).

though strictly speaking we cannot detect this commitment on the premessenger RNA, the contrasting behavior of snoRNAs and U3 snRNAs, compared with other spliceosomal snRNAs, is highly suggestive. Indeed, it has been shown that the elongation rate can affect inclusion of exon E33 of the fibronectin gene without affecting the relative order in which introns are removed (de la Mata et al. 2010). A corollary of this observation is that, in this case, commitment to inclusion can be achieved co-transcriptionally, while actual intron removal might occur later (de la Mata et al. 2010). Hence transcription-mediated influences on splicing are probably larger than can be detected with the data analyzed here.

A variety of recent studies have linked chromatin structure to splicing. Co-transcriptionality of splicing is not an absolute prerequisite for a chromatin-splicing connection, because chromatin

could influence commitment rather than actual intron removal (see above). In the light of our data, it seems, however, that the majority of splicing occurs during transcription and thereby offers an even more direct opportunity for chromatin to influence splicing. Indeed, we detect enrichment of a variety of chromatin marks on exons in the process of being spliced (i.e., exons, with low coSI values in the chromatin-associated RNA).

While proximity to the polyA site seems to disfavor co-transcriptional splicing near the end of the gene, other features such as 5′-to-3′ decreasing intron size and increasing splice site strength favor rapid splicing, so that comparatively high co-transcriptional splicing completion can still be observed toward the 3′ end of the gene. Moreover, it is known that various histone marks vary along the gene (Barski et al. 2007). Such a special chromatin organization toward the 3′ end of the gene could also contribute either directly or indirectly to splicing completion prior to polyadenylation.

Interestingly, gene expression of nuclear polyA+ RNAs is a weak but significant predictor of coSI values in the chromatin total fraction, suggesting a selective pressure for splicing in more highly expressed genes occurs more rapidly. Splicing around coding exons is significantly more often co-transcriptional (in comparison to exons containing noncoding sequence), while splicing around alternatively skipped exons is significantly more post-transcriptional (than for exons not involved in skipping events). Importantly, this does not imply that for all alternative exons splicing of the corresponding introns always occurs post-transcriptionally. Rather, it means that while only a few introns surrounding constitutive exons are removed post-transcriptionally, a significantly higher fraction of introns surrounding (or skipping) alternative exons are removed post-transcriptionally. This genome-wide picture supposedly represents a mixture of two models observed on the fibronectin gene (de la Mata et al. 2010) and on the *Sxl* and *PTBP2* (also known as *nPTB*) genes (Vargas et al. 2011). In the former case, changed exon inclusion levels were achieved without changing the relative timing of actual intron removal but, supposedly, rather by changing splicing commitment co-transcriptionally (de la Mata et al. 2010). In the latter, however, exon inclusion occurred when splicing was co-transcriptional, whereas the exon was skipped when splicing was carried out post-transcriptionally (Vargas et al. 2011). One interpretation for these and our observations is that co-transcriptional splicing tends to be more faithful than postTS, which would therefore offer more opportunities for an exon to be alternatively, that is, differently, included. An interesting corollary of this idea is that when a shorter-than-usual isoform of a gene is expressed, some introns around internal exons might be spliced more often post-transcriptionally, as they are closer to the chosen polyA site. This could then lead to changed inclusion rates of the exon.

Lower coSI values for lncRNAs can be interpreted in multiple ways, all of which probably apply to different subsets of this rather heterogeneous RNA class. Some splicing events in lncRNAs are probably carried out later, that is, post-transcriptionally, simply because lncRNA gene features (e.g., shorter gene length, lower expression) favor this mode of splicing. It is highly likely given the data presented here and previously described examples, that many lncRNAs either (1) remain completely unspliced or (2) have a high proportion of primary transcripts that are never spliced, while a minority are processed by the splicing machinery. For example, two lncRNAs involved in imprinting are likely to remain in the nucleus in an unspliced state: *AIRN* (Sleutels et al. 2002) and *KCNQ1OT1* (Mancini-Dinardo et al. 2006). However, our data should be treated with caution, since the analysis was carried out on the small subset of lncRNAs that are expressed sufficiently highly to calculate a coSI score with confidence (see Supplemental Methods). Nevertheless, the coSI data presented here will be a valuable tool for subclassifying lncRNAs by their processing status.

In summary, we believe that our results strongly suggest that splicing is a highly co-transcriptional process, whose outcome depends crucially on many factors in the exon, and the overall gene sequence, as well as on chromatin architecture and transcription dynamics. As our analysis reveals here, the interrogation of RNA fractions provides invaluable information on the processing pathways establishing RNA genealogy.

Data access

Supplemental Table S1 can be accessed at http://genome.crg.es/~htilgner/2011_coSI_paper/2011cp_index.html. Raw RNA-seq reads can be accessed at the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) under accession numbers GSE30567 and GSE24565. Additional detailed methods for RNA-seq can be obtained in the production documents under "CSHL Long RNA-seq" and "CSHL Sm RNA-seq" at http://genome.ucsc.edu/ENCODE/downloads.html.

Competing interest statement

Michael Snyder is a consultant for Illumina and on the scientific advisory board of Personalis and GenapSys.

Acknowledgments

We thank Juan Valcárcel and Tobias Warnecke from the CRG for useful discussions. This work has been carried out under grants RD07/0067/0012, BIO2006-03380, and CSD2007-00050 from the Spanish Ministry of Science, and grants 1U54HG004557-01 and 1U54HG004555-01 from the National Institutes of Health.

References

Allo M, Buggiano V, Fededa JP, Petrillo E, Schor I, de la Mata M, Agirre E, Plass M, Eyras E, Elela SA, et al. 2009. Control of alternative splicing through siRNA-mediated transcriptional gene silencing. *Nat Struct Mol Biol* **16:** 717–724.

Ameur A, Zaghlool A, Halvardson J, Wetterbom A, Gyllensten U, Cavelier L, Feuk L. 2011. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat Struct Mol Biol* **18:** 1435–1440.

Andersson R, Enroth S, Rada-Iglesias A, Wadelius C, Komorowski J. 2009. Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Res* **19:** 1732–1741.

Barash YCJ, Gao W, Pan Qu, Wang X, Shai O, Blencowe J, Frey B. 2010. Deciphering the splicing code. *Nature* **465:** 53–59.

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129:** 823–837.

Beyer AL, Osheim YN. 1988. Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. *Genes Dev* **2:** 754–765.

Carrillo Oesterreich F, Preibisch S, Neugebauer KM. 2010. Global analysis of nascent RNA reveals transcriptional pausing in terminal exons. *Mol Cell* **40:** 571–581.

Cramer P, Pesce CG, Baralle FE, Kornblihtt AR. 1997. Functional association between promoter structure and transcript alternative splicing. *Proc Natl Acad Sci* **94:** 11456–11460.

de la Mata M, Alonso CR, Kadener S, Fededa JP, Blaustein M, Pelisch F, Cramer P, Bentley D, Kornblihtt AR. 2003. A slow RNA polymerase II affects alternative splicing in vivo. *Mol Cell* **12:** 525–532.

de la Mata M, Lafaille C, Kornblihtt AR. 2010. First come, first served revisited: Factors affecting the same alternative splicing event have different effects on the relative rates of intron removal. *RNA* **16:** 904–912.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res* (this issue). doi: 10.1101/gr.132159.111.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi AM, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. 2012. Landscape of transcription in human cells. *Nature* (in press).

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473:** 43–49.

Graveley BR. 2000. Sorting out the complexity of SR protein functions. *RNA* **6:** 1197–1211.

Hon G, Wang W, Ren B. 2009. Discovery and annotation of functional chromatin signatures in the human genome. *PLoS Comput Biol* **5:** e1000566. doi: 10.1371/journal.pcbi.1000566.

Howe KJ, Kane CM, Ares M Jr. 2003. Perturbation of transcription elongation influences the fidelity of internal exon inclusion in *Saccharomyces cerevisiae*. *RNA* **9:** 993–1006.

Kadener S, Cramer P, Nogues G, Cazalla D, de la Mata M, Fededa JP, Werbajh SE, Srebrow A, Kornblihtt AR. 2001. Antagonistic effects of T-Ag and VP16 reveal a role for RNA pol II elongation on alternative splicing. *EMBO J* **20:** 5759–5768.

Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahringer J. 2009. Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat Genet* **41:** 376–381.

Kundaje A, Kyriazopoulou-Panagiotopoulou S, Libbrecht M, Smith CL, Raha D, Winters EE, Johnson SM, Snyder M, Batzoglou S, Sidow A. 2012. Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Res* (this issue). doi: 10.1101/gr.136366.111.

Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T. 2010. Regulation of alternative splicing by histone modifications. *Science* **327:** 996–1000.

Mancini-Dinardo D, Steele SJ, Levorse JM, Ingram RS, Tilghman SM. 2006. Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. *Genes Dev* **20:** 1268–1282.

Nahkuri S, Taft RJ, Mattick JS. 2009. Nucleosomes are preferentially positioned at exons in somatic and sperm cells. *Cell Cycle* **8:** 3420–3424.

Nogues G, Kadener S, Cramer P, Bentley D, Kornblihtt AR. 2002. Transcriptional activators differ in their abilities to control alternative splicing. *J Biol Chem* **277:** 43110–43114.

O'Mullane L, Eperon IC. 1998. The pre-mRNA 5′ cap determines whether U6 small nuclear RNA succeeds U1 small nuclear ribonucleoprotein particle at 5′ splice sites. *Mol Cell Biol* **18:** 7510–7520.

Pandya-Jones A, Black DL. 2009. Co-transcriptional splicing of constitutive and alternative exons. *RNA* **15:** 1896–1908.

Roberts GC, Gooding C, Mak HY, Proudfoot NJ, Smith CW. 1998. Co-transcriptional commitment to alternative splice site selection. *Nucleic Acids Res* **26:** 5568–5572.

Schor IE, Rascovan N, Pelisch F, Allo M, Kornblihtt AR. 2009. Neuronal cell depolarization induces intragenic chromatin modifications affecting NCAM alternative splicing. *Proc Natl Acad Sci* **106:** 4325–4330.

Schwartz S, Meshorer E, Ast G. 2009. Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol* **16:** 990–995.

Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, Oberdoerffer P, Sandberg R, Oberdoerffer S. 2011. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479:** 74–79.

Sleutels F, Zwart R, Barlow DP. 2002. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* **415:** 810–813.

Smith CW, Valcárcel J. 2000. Alternative pre-mRNA splicing: The logic of combinatorial control. *Trends Biochem Sci* **25:** 381–388.

Spies N, Nielsen CB, Padgett RA, Burge CB. 2009. Biased chromatin signatures around polyadenylation sites and exons. *Mol Cell* **36:** 245–254.

Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, Valcarcel J, Guigo R. 2009. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol* **16:** 996–1001.

Vargas DY, Shah K, Batish M, Levandoski M, Sinha S, Marras SA, Schedl P, Tyagi S. 2011. Single-molecule imaging of transcriptionally coupled and uncoupled splicing. *Cell* **147:** 1054–1065.

Wang Z, Burge CB. 2008. Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA* **14:** 802–813.

Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. 2004. Systematic identification and analysis of exonic splicing silencers. *Cell* **119:** 831–845.

Received November 6, 2011; accepted in revised form February 7, 2012.

Results – Part III

Promoter-like epigenetic signatures in exons with cell type-specific splicing

Joao Curado[1,2,*], Camilla Iannone[1,3*], Hagen Tilgner[1,4,*], Juan Valcárcel[1,3,5], Roderic Guigó[1,3,+]

[1] Centre for Genomic Regulation (CRG) and UPF, Dr. Aiguader, 88, 08003 Barcelona, Spain

[2] Graduate program in Areas of Basic and Applied Biology, Abel Salazar Biomedical Sciences Institute, University of Porto, 4099-003 Porto, Portugal

[3] Universitat Pompeu Fabra, Dr. Aiguader 88, 08003 Barcelona, Spain.

[4] Department of Genetics, Stanford University, 300 Pasteur Dr., Stanford, CA 94305-5120, USA.

[5] Institució Catalana de Recerca i Estudis Avançats, Pg Lluis Companys 23, 08010 Barcelona, Spain.

* equal contribution

+ To whom correspondence should be addressed

Abstract

Background

Pre-mRNA splicing occurs, in large part, co-transcriptionally and both nucleosome density and histone modifications have been proposed to play a role in splice site recognition and regulation. The extent of these interplay, and the mechanisms underlying it, remain however, poorly understood.

Results

We used transcriptomic and epigenomic data generated by the ENCODE project to investigate the association between chromatin structure and alternative splicing. We did found a strong and significant positive association between H3K9ac, H3K27ac, H3K4me3 and inclusion in a small, but well defined class of exons (about 4% of all regulated exons). These exons are systematically maintained at comparatively low levels of inclusion across cell types, but their inclusion is significantly enhanced when in physical proximity (either in linear or in the 3D space) of active promoters.

Conclusion

Histone modifications and other chromatin features that activate transcription could be co-opted to participate in the regulation of the splicing of exons that are in the physical proximity of promoter regions.

Key words

Chromatin; splicing; histone; modifications; promoters; proximity; transcription; epigenetic; RNA; ENCODE;

Introduction

Alternative pre-mRNA splicing is assumed to expand the diversity of mRNAs encoded in the genome. The prevalence of alternative splicing increases from invertebrates to vertebrates [1] and is particularly high in the immune and nervous systems, where high diversity of molecular repertoires is necessary for cell identity [2]. Whether an alternative exon is included or excluded in a mature RNA is considered a matter of combinatorial control, involving splice sites, additional binding sites and the factors that recognize them [3]. Recent evidence suggests that chromatin organization and transcriptional dynamics may also contribute to this control. First, splicing can occur co-transcriptionally [4, 5], and this has been demonstrated to be widespread in yeast [6], fruit fly [7], and human [8, 9]. Second, some splicing factors are known to interact with modified histone tails, and intragenic histone modifications have been shown to be involved in alternative splicing decisions on individual genes [10]. Third, RNA Polymerase II elongation dynamics is known to influence exon inclusion [11, 12], which was shown to be modulated by CTCF binding [13]. Lastly, a number of independent studies demonstrated that nucleosome density correlates with exon-intron architecture genome-wide [9, 14-19], and specifically with exon inclusion levels [20, 21].

While links between chromatin and splicing have thus been established, attempts to incorporate chromatin information on quantitative models predictive of cell type specific exon inclusion levels have met so far with moderate success [22], and the extent to which cell-type specific

chromatin organization contributes to cell-type specific splicing patterns remain largely unknown.

Here to investigate the relationship between chromatin and splicing, we analyzed transcriptome and epigenome data generated in a number of human cell lines within the ENCODE project [23, 24]. First, we used RNASeq data to identify exons that are differentially included between cell lines. We found that a relatively small fraction of human exons (about 3% of all internal exons) exhibit regulated inclusion across human cell lines. These regulated exons are maintained at intermediate inclusion levels compared to all exons. Second, we used ChIPSeq data to investigate the association between the inclusion of regulated exons and histone modifications. Our results strongly suggest that there is little or no direct association between histone modfications and the inclusion levels of the majority of these exons.. We identified, however, a small set of regulated exons (about 4% of all differentialy included exons) in which cell type specific inclusion levels do appear to be directly associated to levels of canonically activating histone modifications. In contrast to most exons, the inclusion of these exons is maintained at remarkabe low levels across a large variety of cell types and tissues. In addition of being enriched in histone modifications, these exons have other characteristics typical of promoter regions, but they do not correspond to sites of transcription initiation. However, they tend to lay closer to transcription initiation sites, and through chromatin looping they tend to interact with promoter regions.

Our observations are consistent with a role for promoter regions and for promoter-characteristic epigenetic signatures in the regulation of the alternative splicing of a well-defined set of exons, possibly involving the opening of chromatin and folding of chromatin loops that bring together regulated exons and promoters into close spatial distance. Histone modifications and other features that activate transcription could then be co-opted in these cases to participate also in the regulation of exon inclusion.

Results

R1. Alternatively included exons in pair-wise cell type comparisons

We used nuclear polyA+ RNASeq samples from five Tier 1 human cell lines from the ENCODE project (K562, Gm12878, Hepg2, Huvec, Helas3) [25] to identify differentially included internal exons in pair-wise comparisons of cell lines. We used a method similar to that published by Wang and coworkers [26] (Figure 1a, Methods, Figure S1). Nuclear polyA+ RNA was selected, since, in contrast to other RNA fractions, in this fraction splicing has been essentially completed [9], but it is unlikely to have undergone nonsense mediated decay (NMD), and thus, it reflects more precisely the direct outcome of splicing.

We selected 73,329 internal exons with canonical splice junctions that lay at least 600bp away from the closely annotated Transcription Start Site (TSS) and Transcription Termination Site (TTS), and for which there were enough RNASeq reads to compute differential inclusion in at least one pairwise cell comparison (see Methods). We used the one-side Fisher test on the number of exclusion and inclusion reads to identify differentially included exons between each cell-pair (see Methods). Exons with significant inclusion level changes were called more/less included exons depending in the direction of the change (Figure 1a). Figure 1b-g shows the results of the comparison of Gm12878 and K562. We identified 1688 exons regulated between these two cell lines (1066 more included in Gm12878 and 622 more included in K562, p-value <0.05, Figure 1b).

These differentially included exons possess known properties of alternative exons, as previously described in the literature [27, 28]: 1) they have weaker splice sites compared to exons not differentially included (Figure 1c), 2) they tend to be shorter (Figure 1d) and, 3) when coding, their length is more often divisible by three (Figure 1e). Moreover, 4) we did not find significant differences in the expression levels of the genes hosting differentially included exons between the two cell lines compared (Figure 1f). In order to independently validate our alternative exon calling method, we selected a total of 15 of these exons (Figure 1g). Using exon-junction oligonucleotides (Table S1), we quantified by qPCR the ratio of inclusion/skipping isoform in Gm12878 compared to K562. This method of quantification provides an assessment of the differences in the relative inclusion of the exons regardless of possible differences in gene expression between the cell lines compared. We validated 12 out of the 15 selected cases (Figure 1g), corresponding to a validation rate of 80%.

We assessed whether conditions 1-4 above (Figure 1c-f) where satisfied in each of the ten pairwise comparisons between the five cell lines considered here, and we kept only the seven comparison satisfying all of them (Table S2). Furthermore, we retained only exons with minimum absolute change of 0.1 or 2-fold in the inclusion levels between the two cell lines. In total, we obtained 1849 more included and 2483 less included exon comparisons in the seven cell-pairs employed. The terms “more included exons” and “less included exons” are arbitrary, since they depend on the direction of the comparison. We preferred to keep them

separate to allow for better visualization and validation of the results presented (see below).

Because the same exon can appear in different pairwise comparisons, when pooled together, the two sets correspond to 2081 unique exons that showed regulated inclusion levels across the human ENCODE cell lines. This corresponds to about 3% of all exons initially considered. Exons with regulated inclusion exhibit in general weak inclusion changes across human cell lines (Figure 2a). The median exon inclusion range (that is, the differences between the maximum and minimum inclusion observed) is 0.20. For more than 90% of the exons, the change is less than 0.5. Moreover, they show generally intermediate exon inclusion levels when compared with the inclusion levels of non-regulated exons (Figure 2b).

Because gene expression levels are linked to chromatin organization [29] and can also influence splicing [3], we excluded exons from genes showing large (more than 10 fold) expression differences between cell lines. In addition, in our analysis, we used only one exon per gene, the one with the lowest p-value. In total, we obtained 1684 more included and 2198 less included exon comparisons in the seven cell-pairs employed corresponding to 1921 unique exons (Table S3, Additional file 1 and 2).

R2. Co-occurrence of differences in histone modifications with alternative exon inclusion

For each differentially included exon in each pairwise cell comparison, we computed differential signal for nine histone marks, the

insulator protein CTCF and input DNA [24, 30, 31]. We defined the "differential signal" for each chromatin feature as the difference of the average normalized signal over the exon in the second cell type (e.g. Gm12878 for the comparison K562 vs. Gm12878) and the average signal over the exon in the first one (e.g. K562 for the same comparison, see Methods). We pooled the differential exonic signal across all exons and all comparisons together to produce a single composite comparison for each monitored variable separately for "more included" and "less included" exons. By using differential signals on the same genomic interval, we eliminated possible biases due to intrinsic genomic features such as GC content.

Our analysis revealed enrichment of CTCF, H3K9ac, H3K27ac and H3K4me3 levels in more included exons (p-value< 0.01; Wilcoxon signed-rank test with Bonferroni correction) (Figure 3a). We also observed a negative association with the control signal, which consists of cross-linked and sonicated DNA. This may represent a measure of chromatin compaction, and may reflect an association between open chromatin and higher levels of exon inclusion. Importantly, the association observed between exon inclusion and input DNA is in the opposite direction than that observed for rest of the chromatin features, indicating that we are likely underestimating the strength of the associations. To validate these results, we performed ChIP-qPCR using H3K9ac antibodies and primers specific for the target exons (and a constitutive exon of the same gene as a control) in K562 and Gm12878 cells. In all four alternative exons investigated, a clear difference in H3K9ac signal between the two cell lines

was detected, positively correlating with differential exon usage. In contrast, constitutive exons on the same genes showed, in general, smaller (sometimes even opposite) differences and higher variability among replicates (Figure 3b).

To assess whether the differential enrichment in histone modifications was local and specific of the regulated exons, or rather affected more extensive regions of the gene, we analyzed the distribution of epigenetic marks in the closest upstream and downstream exons that our method did not identify as differentially included (Methods). This defines a "not regulated -regulated - not regulated" exon triplet. Differential chromatin profiles were calculated within 800bp windows from the center of the exons. The results showed that the enrichment in chromatin signals does not extend to the flanking exons and it is, therefore, specific of the differentially included exons (Figures 3c-f). This confirms a significant positive association between exon inclusion and local levels of H3K9ac, H3K27ac, H3K4me3 and CTCF binding.

R3. Promoter-like histone marks and exon inclusion

While the results above do indicate a significant association between a number of histone modifications and exon inclusion, the effect is certainly weak. This could reflect a general, but weak effect of histone modification on most differentially included exons, or alternatively, a strong effect only on a subset of them. To investigate the two alternatives, we performed k-means clustering on the sets of more and less included

exons, based on the levels of the five chromatin features that we found significantly associated with differential exon inclusion. After k-means optimization, the data was partitioned in four clusters of different signal profiles and each exon was assigned to the cluster with the nearest mean (Figure 4). As expected “more included” and “less included” exons generated similar but mirrored clusters.

While the majority of differentially included exons do not show differences in the levels of the monitored histone modifications, a subset of exon comparisons (59 “more included” and 41 “less included”, corresponding to 70 unique exons) exhibits large differential levels of H3K9ac, H3K27ac and H3K4me3 associated with differential exon inclusion (Figure 4, Additional file 3). Thus, while the inclusion of the majority of exons does not appear to be directly associated to levels of chromatin marks, some histone modifications are strongly correlated with the regulation of the inclusion of a subset of exons (about 4% of all differentially included exons). Principal Component Analysis (PCA) confirmed these results (Figure S3).

To further validate these findings, we compared K562 and NHEK, an ENCODE Tier 2 cell line that had not been previously used in our analysis. Out of the 70 exons above, 22 are differentially included between these two cell lines. In 17 of them (77%), the direction of the differential inclusion is consistent with the direction of the differential levels of H3K9ac, H3K27ac and H3K4me3 (Figure S3), while only in three cases (14%) exon inclusion and histone modification levels change in opposite

directions. In the remaining two cases, no change in histone modification levels could be detected.

These exons, which inclusion levels appear to be associated with levels of H3K9ac, H3K27ac and H3K4me3, will be referred to as “promoter-like” exons since these histone modifications are known signatures of promoters. To further characterize them, we merged the two sets of “promoter-like exons” in one single group, retaining the association with the pair of cell lines where they were identified as differentially included. From now on, these cell lines will be referred to as “C-higher” (the cell line in which the exons are more included) and “C-lower” (the cell line in which they are less included).

Compared to the rest of regulated exons, “promoter-like” exons are included at particularly low levels (median inclusion level of 0.32 and 0.07 in “C-higher” and “C-lower” cell lines, respectively, compared with 0.54 and 0.23 of the non promoter-like regulated exons (Figure 5a). To assess whether low inclusion levels are constitutive of “promoter-like” exons, or a consequence of our measurements restricted to human cell lines, we used 1,500 RNASeq samples from the GTEx project [32], to estimate exon inclusion levels of the set of regulated exons in human tissues. We found that also in human tissues, “promoter-like” exons exhibit significantly lower levels of inclusion that regulated “non promoter-like” exons (Fig S4).

“Promoter-like” exons are characterized by additional promoter-associated features when compared to the rest of regulated exons. First, they are enriched in binding sites, both when considering sequence motifs (Table S5 and S6), and accumulation of Transcription Factor (TF) ChIPSeq

reads. Indeed, we found that 15 out of the 32 TF analyzed have significantly more accumulation of reads in C-higher than in C-lower cell lines, while none displays the opposite trend (Figure 5b and S5). Among the enriched transcription factors is Brg1, which together with Brm is one of the two ATPases of the chromatin remodeling complex SWI/SNF, and it has been shown to interact with the splicing machinery [33] (Figure S5b). Second, they are enriched in DNase I hypersensitive sites (DHS) in C-higher cell lines (Figure 5c). The difference in DHS signal is present in the regulated exon but not in the surrounding non-regulated ones. DHS are indicative of open, more accessible chromatin; they usually mark *cis*-regulatory elements, including promoter and enhancer sequences. We also found a weaker enrichment in the set of all regulated exons (Figure S6). Third, under conditions of higher inclusion, they also show a clear enrichment of RNA Polymerase II signal (Figure 5d).

All these promoter associated features in "promoter-like exons" could suggest that these exons actually overlap un-annotated TSS—in which case, elevated levels of H3K9ac, H3K27ac and H3K4me3 could simply reflect the action of transcription, and be unrelated to exon inclusion. To rule out this possibility, we analyzed the number of CAGE tags (sequence tags that target specifically the 5' end of transcripts [34] mapping to "promoter-like" exons, and find it marginal when compared to the level in annotated TSS (Figure 5e). We also found no significant difference in the distribution of upstream and downstream junction inclusion reads (RNASeq reads that connect two neighboring exons),

further confirming that “promoter-like exons” are indeed “bona fide” exons, and do not represent, as a bulk, un-annotated TSSs (Figure S7).

R4. Proximity to the promoter and exon inclusion

In spite of not being promoters themselves, “Promoter-like” exons are significantly closer to the annotated TSSs when compared to the rest of differentially included exons (Figure 6a), and actually most of them are second exons of the transcript, with their upstream exon beginning at the TSS itself. This explains the enrichment of DHS and Pol II signals in the non-regulated exons upstream of “promoter-like exons” (Figures 5c,d)—albeit these enrichments are not significantly different between C-higher and C-lower cell lines, in contrast to the enrichments in the “promoter-like” exons. Moreover, analyzing clusters of CAGE tags, which are assumed to indicate annotated or un-annotated TSS, we observed alternative TSS usage for “promoter-like” exons in C-higher and C-lower conditions (Figure 6b). We found that in 45 of the 100 “promoter-like” exon comparisons, the active TSS closest to the exon is closer in the C-higher cell line than in the C-lower cell line, while in only 6 exon comparisons is the other way around (Figure 6b,c). For eight exon comparisons we found CAGE clusters within the exonic region, suggesting that these exons could indeed correspond to new, un-annotated TSSs. This could explain the significant CAGE enrichment occupancy in “promoter-like” exons when comparing C-higher and C-lower cell lines (Figure 5e). Figure 6d shows an example of an exon with higher inclusion level in Hela than in Gm1278 and with a closer active TSS being used only in the cell line of higher inclusion.

These results suggest that regulated enrichment of histone modifications and other promoter-associated features in “promoter-like” exons could be the consequence of the physical proximity between these exons and real promoters. Transcription activating histone modifications would thus be “co-opted” to participate also in the regulation of inclusion of “promoter-like” exons. To further test this hypothesis, we analyzed genome-wide Chromatin Interaction Analysis with Paired-End-Tag sequencing (ChIA-PET) data that has been used to map long-range chromatin interactions associated with RNA polymerase II un-phosphorylated ser2, found in the transcription pre-initiation complex, a characteristic mark of the gene promoters [35]. Indeed, it has been recently shown that some internal exons loop and physically interact with promoter and enhancers, and for this reason display “promoter-like” or “enhancer-like” chromatin marks, and are enriched for co-transcriptional splicing [36]. In “promoter-like” exons we found enriched ChIA-PET signal (indicative of an increase in looping and interaction with the promoter), in C-higher than C-lower conditions (Figures 6e). As a control, we did not find ChIA-PET signal enrichment in the set of all regulated exons (Figure S8). Figure 6f shows a “promoter-like” exon more included in Hela than in K562, exhibiting local peaks of DNase I, RNA Pol II, H3K9ac, H3K27ac and H3K4me3 in Hela, but not in K562. The exon also shows ChIA-PET tags only in Hela cells.

Discussion

Regulated alternative splicing is assumed to contribute to cell type identity and methods have been developed which are able to predict tissue specific exon inclusion with high accuracy [37]. In our analysis, we found a relatively small number of human exons (about 3% of all exons) exhibiting regulated inclusion in a panel of human cell lines. This cannot be attributed to insufficient sampling by RNASeq, since ENCODE cell lines are sequenced in replicates at very high depth of coverage (around 240M reads per sample). On the other hand, the diversity of biological samples used is certainly reduced, and cell lines are known to exhibit peculiar biology [38]. While regulated splicing, therefore, is likely to be more widespread than detected here, our results could also suggest that the contribution of splicing regulation to defining cell type identity is exerted chiefly through a relatively small, but well defined, set of exons.

Recent results have unveiled that pre-mRNA splicing occurs predominantly co-transcriptionally, thus providing a framework in which chromatin and transcription-related factors interact with the pre-mRNA processing machinery. However, among most exons with regulated inclusion we found in general, little direct association between differential inclusions and histone modifications. While these results are not fully unexpected, since splicing factors are likely to be the main players in splicing regulation, they somehow in contrast with reports of histone modifications influencing splicing outcomes through recruitment of splicing factors and through the modulation of RNA Pol II dynamics.

Indeed, previous work linked high H3K9ac levels in the *NCAM* gene with fast elongating RNA Pol II and skipping of a specific exon [39]. High levels of H3K36me3 or H3K27me3 along the *FGFR2* were correlated with the regulation of a mutually exclusive alternative splicing event [10]. In these cases, however, changes in histone modifications appear to spread over large regions covering the whole gene, while here we explicitly explored chromatin modifications local to the exons. More importantly, in our work we investigated only direct effects acting independently, and we ignored the role of high order interactions between different histone modifications and other elements of chromatin structure. These could actually configure a quite complex histone-based splicing regulatory code. Furthermore, we focused specifically in complete exon skipping events, and ignored other types of splicing events such as alternative splice site usage. In this regard, Tilgner et al. [21] found that nucleosome occupancy may contribute more strongly to the definition to the 3' splice site. If so, histone modifications would also be expected to play a major role in the regulation of alternative 3' splice sites.

While we did not found evidence for a general direct effect of chromatin structure in exon inclusion, we did identify a subgroup of regulated exons (about 4% of all regulated exons) for which co-occurrence of H3K9ac, H3K27ac and H3K4me3, histone modifications typically associated to promoters, strongly correlate with exon usage levels. This association is not biased by our discovery approach, since we replicated it in cell line comparisons that were not part of the training set. These "promoter-like" alternative exons appear predominantly in low abundance

isoforms, but in which, a significant increase in the density of histone modification correlates with an increase in the levels of exon inclusion. These observations suggest that chromatin architecture may play a more prominent role in the regulation of exon inclusion, under conditions of weak splice site recognition.

We further related the accumulation of these histone modifications in highly included exons with higher occupancy of RNA Polymerase II. Accumulation of RNA Polymerase II has been linked to exon inclusion [40, 41], associated with slower Pol II kinetics and consequently additional opportunities for splice site recognition before competing sites come into play [42]. However we also found higher inclusion of “promoter-like” exons in states of open chromatin, as measured by DNase I. This observation is somehow in contrast with previously proposed models linking closed chromatin and slower transcription elongation with increased exon inclusion [11, 43-45]. In particular for H3K9ac, previous studies reported a correlation between accumulation of this mark along the whole gene body (NCAM) and skipping of a specific alternative exon [45].

We also found that “promoter-like” exons, while no promoters themselves, they are, on average, closer to TSS and enriched by ChIA-PET tags associated with RNA Polymerase II. Thus, we hypothesize that, in these exons, splicing regulation is mediated by the promoter, either by formation of a DNA loop with the exon and helping chromatin marking and transcription factor binding extending from the TSS to the alternative exon, or by differential splice site pairing when an alternative TSS generates an alternative first exon (Figure 7). Recent work exploring the

three-dimensional structure of the genome reported physical links between internal exons and their associated promoter or enhancers. These results argue for an interplay between 3D-genome organization and alternative splicing regulation and warrant the systematic analysis of these associations in future studies using conformation capture technologies [46]. An alternative interpretation relies on the fact that H3K9ac, H3K27ac, H3K4me3, and DNase hypersensitive on “promoter-like” exons simply reflect open chromatin on these exons. It is conceivable that binding of factors facilitated by the opening of chromatin influence splice site recognition either directly through their effects on splicing factor recruitment or through effects on RNA Pol II elongation - a mechanism resembling promotion of exon inclusion by CTCF [13].

In summary, our work sheds light on functional connections between chromatin structure and pre-mRNA processing, establishing associations between epigenetic marks and differential exon inclusion and suggesting a role for promoter-like regions and 3-dimensional genome architecture in the regulation of the alternative splicing of certain exons. We specifically propose that in exons that are proximal to active promoter regions (either in linear or tri-dimensional space), open chromatin promotes exon inclusion, maybe by facilitating the recruitment of splicing factors. However, we want to stress that through our analysis we are unable to uncover the direction of the causation, and while histone modifications have been proposed to promote splicing, results have also been obtained suggesting that splicing can promote modification of histones by enhancing the recruitment of chromatin remodeling factors

[47]. Further research will be needed to work out their detailed molecular mechanisms behind these observations.

Methods

Alternatively skipped exon calling

Using the gencode [48] v15 annotation we determined all exons that are

1. internal in all transcripts they appeared in
2. not overlapped by any non-identical exon in both annotations. Identity of exons is defined by their location (chromosome, start, end strand)
3. between 50 and 450bps long
4. at least 600nts away from the respective annotated TSS or TTS
5. surrounded by AG-GT splice sites
6. located on chromosomes 1-22 and X

For the remaining exons, a two by two table was constructed containing junction inclusion reads and junction exclusion reads in the two cell types (cell1 and cell2) retaining only the exons with a minimum of 1 junction inclusion read in cell1 and 1 exclusion read in cell2 or vice-versa. For every cell-pair, two one-sided Fisher tests were run and corrected for multiple testing in the Benjamini-Hochberg sense, resulting in three disjoint sets of exons:

1. exons that are significantly more included in cell1 (which will be referred to as “more included”, even though the choice of the direction from cell1 to cell2 is clearly arbitrary)
2. exons that are significantly less included in cell1 (which will be referred to as “less included”)

3. exons whose inclusion is not significantly changed between the two cell types (which will be referred to as “notAS exons” for the sake of conciseness and clarity, although “non-significant AS exons” would be more correct.)

From the set of more and less included exons, we further selected the exons that met the following criteria:

i. The expression of the gene containing the exon did not change more than 10fold between cell1 and cell2. To measure gene expression, we used CAGE tags mapping to the gene promoter (see below).
ii. at least 75% of all positions in a 900bp window around the acceptor were uniquely mappable for 36mers (see below).
iii. The inclusion levels of the exon changed by at least 0.1 or two-fold between the two cell lines.

Genes frequently contained more than one alternatively spliced exon thus defined. In order to avoid gene specific bias that might be introduced by genes that contribute many alternative exons (as for example, the TTN gene where 212 exons passed the Fisher test), we chose only one up-regulated and up to one down-regulated exon per gene: The exon with the lowest p-value among all exons for the gene in question.

For non-AS exons a similar procedure was carried out, removing however the “inclusion changed by at least 0.1 or two-fold”-criterion and choosing the exon per gene whose estimated inclusion change was

minimal among all non-AS exons of that gene (instead of the exon with the smallest p-value). Figure S1 illustrates this approach.

Exon triplets

For each regulated exon and each cell type comparison, we defined two non-regulated exons: The closest up- and downstream exon that

- appeared in a transcript together with the alternative exon
- that showed Benjamini-Hochberg-corrected p-value of 0.05 or greater.

Inclusion level calculation

Inclusion level (IncLevel) is a measure defined to describe the splicing status of the exons. It is computed as a function of the reads arguing for the inclusion of the exon (JIR) and reads arguing for exclusion (JER). Formally it is

$$IncLevel = \frac{0.5 * JIR}{0.5 * JIR + JER}$$

A value of 0 represents a totally excluded exon, while a value of 1 represents a totally included exon. IDR (Irreproducible Discovery Rate), a measure widely used within the ENCODE project to assess reproducibility between replicates [49], was applied at a level of 0.01 and only the exons passing this filtered were used in the remaining analysis.

Splice site strength measure

For each exon we used maxEnt [50]in order to calculate an acceptor score and a donor score and represented the "exon strength" by the sum of these two scores.

Gene expression calculation

We employed ENCODE provided CAGE-clusters filtered by an Hidden Markov Model algorithm to differentiate between 5' capped termini of Pol II transcripts and recapping events, and scored according to number of constituent CAGE tags [25].

We associated CAGE-clusters to the closest TSS within a radius of 100 nucleotides. We computed the expression of a given gene as the sum of the scores of CAGE-clusters associated to all gene's TSS.

Mappability calculation

Mappability for the hg19 genome was calculated using the GEM-mapper for 36bp and 75bp reads. For each acceptor in the genome, mappability was calculated in the direction of transcription.

Exon selection for validation experiments

Out of the exons that had cell type specific H3K9ac peaks that co-occurred with high exon inclusion, we selected a total of 12 exons for validation by RT-PCR and ChIP.

Cell culture, RNA extraction and RT-PCR analysis

K562 and Hela cells were grown in Dulbecco's modified Eagle's medium (Gibco BRL) supplemented with 10% of fetal bovine serum (FBS), penicillin and streptomycin. Gm12878 cells were grown in RPMI (INSERT REF), supplemented with 15%FBS (Gibco BRL), glycine (SEE), penicillin and streptomycin.

Total RNA was isolated using Qiagen RNeasy mini kit and re-suspended in RNAse-free water (Ambion). DNA digestion was performed using RNase-free DNase (Promega). DNA-free total RNA (1 μg) was used for RT–PCR using SuperScript III reverse transcriptase (Invitrogen), random hexamers and oligo dT. 5% of the reaction was used for Real Time PCR (Applied Biosystem) together with the primers (Table S2) following the manufacturer's instructions.

Chromatin immunoprecipitation (ChIP).

Cells were plated at a density of $2X10^5$ cells in 75 cm^2 flasks and after 48h of culture, incubated with 1% (vol/vol) formaldehyde in culture medium for 10 minutes at room temperature. Cells were then washed in cold phosphate-buffered saline (PBS), harvested and lysed in a buffer containing 1% SDS, 10 mM EDTA and 50 mM Tris/HCl pH 8.1, and sonicated in 15ml tubes with Bioruptor UCD-200 Diagenode (ultrasonic wave output power 250W, 30" on-30"off, 4X10') to yield chromatin sizes of 150-300 bp. 100 μg of DNA/sample were used for immunoprecipitation with 5μg of anti-H3K9ac rabbit (ab4441), anti-H3 rabbit (ab1791) or control rabbit IgGs (Sigma-Aldrich). Co-precipitated DNA was then

analyzed by Quantitative real time PCR performed with Sybr Green mix (Applied Biosystem) according to manufacturer's instructions. The antibody against total H3 was used for normalization as well as a control to exclude the possibility that the effects observed are caused by differences in nucleosome occupancy. The primers used are listed in the Supplementary Table 1.

Chromatin signature in exons

To define the chromatin signature of each exon, we calculated the average of signal, per histone modification, from ENCODE normalized signal files in the exon region [51]. The same procedure was applied for the control, DNase I, ChIA-PET and transcription factors tracks. Wilcoxon signed-rank test, with Bonferroni correction for multiple testing, was used to assess the significance between the groups of exons.

Exon clustering

Alternative spliced exons differentially expressed between cell lines were partitioned according to their histone modifications and control levels using a k-means clustering approach. The method was applied to "higher included" and "lower included" exons separately. Exons falling in cluster 3 and 4 of "higher included" exons and in cluster 4 of "lower included" exons were merged in the group of "Promoter-like exons". Each exon was thus associated with 2 cell lines, one in the "higher inclusion" group and one in the "lower inclusion" group.

Distance to closest annotated TSS

The distance of an exon to the annotated TSS was calculated by measuring the genomic distance from the first nucleotide of the exon and the closest TSS from all the transcripts the exon belongs to.

Distance to closest used TSS

The distance of an exon to the closest used TSS was calculated by finding the closest CAGE-cluster with a minimum expression value of 1 and calculating its genomic distance to the studied exon, in an annotation independent manner.

Binding motifs analysis

An hyper geometric test, with Benjamin Hochberg p-value correction, was applied to JASPAR CORE 2014[52] and MEME 4.4[53] databases looking for enrichment in binding transcription factor and RNA binding proteins motifs inside the exons respectively, in the “Promoter-like exons” versus the remaining differentially included identified exons.

Splicing change in NHEK

NHEK, a cell line not used in the discovery analysis of the paper, was used for validation of our findings. The difference in inclusion and in histone modification signal was calculated between NHEK and K562, Only exons with an inclusion difference larger than 0.1 were used. Total signal was calculated as the sum of H3K27ac, H3K4me3 and H3K9ac differences.

Only exons with a Total difference higher than 1 were considered as having chromatin differences.

Exon inclusion in GTEx project

We calculate inclusion levels, as described above, for the set of all internal exons in the 1,493 post-mortem samples available from the GTEx project. The samples are very heterogeneous, coming from up to 43 different tissues from 175 individuals

List of abbreviations

bp: Base pairs

CAGE: Cap analysis gene expression

ChIA-PET: Chromatin Interaction Analysis with Paired-End-Tag sequencing

ChIP: chromatin immunoprecipitation

ChIPSeq: Sequencing of chromatin immunoprecipitation

CTCF: CCCTC-binding factor

DHS: DNase I hypersensitivity sites

DNA: Deoxyribonucleic acid

ENCODE: Encyclopedia of DNA Elements

GTEx: Genotype-Tissue Expression project

H3K9ac: H3 lysine 9 acetylation

H3K27ac: H3 lysine 27 acetylation

H3K4me3: H3 lysine 3 trimethylation

IDR: Irreproducible discovery rate

JIR: Junction inclusion read

JER: Junction exclusion read

mRNA: Messenger ribonucleic acid

NMD: Nonsense mediated decay

Pol: Polymerase

PolyA+: Polyadenylated

PCA: Principal component analysis

qPCR: quantitative real-time polymerase chain reaction

RNA: Ribonucleic acid

RNASeq: Sequencing of RNA

RT-PCR: Reverse transcription polymerase chain reaction

TF: Transcription factor

TSS: Transcription start site

TTS: Transcription termination site

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All the authors conceived the study and wrote the paper. JC and HT carried out the data analysis. CI performed the validation experiments. All authors read and approved the final manuscript.

Description of additional data files

The following additional data files are available with the online version of the paper. Additional files 1 and 2 contain the set of differentially included exons selected in all cell-pairs used, more and less included exons, respectively. These files are in bed format and field 4 is the cell-pair in which the exon was identified as regulated. Additional file 3 contains the list of “promoter-like”, also in bed format. Additional file 4 contains the supplementary figures and tables.

Acknowledgements

Research reported in this publication was supported by the NHGRI award 1U54HG007004, the Spanish Ministerio de Economia y Conocimiento (MINECO) grant BIO2011-26205 and the ERC/European Community PF7 grant 294653 RNA-MAPS. JC was supported by a SFRH/BD/33535/2008 from the Portuguese Foundation to Science and Technology. CI was supported by a La Caixa predoctoral fellowship. Work in JV’s lab was supported by Fundación Botín, by Banco de Santander through its Santander Universities Global Division and by Consolider RNAREG, MINECO and AGAUR. We thank Anshul Kundaje, Ben Brown, Michael Snyder and Thomas Gingeras for useful discussions and access to data.

References

1. Kim E, Magen A, Ast G: **Different levels of alternative splicing among eukaryotes.** *Nucleic Acids Res* 2007, **35:**125-131.
2. Nilsen TW, Graveley BR: **Expansion of the eukaryotic proteome by alternative splicing.** *Nature* 2010, **463:**457-463.
3. Fu XD, Ares M, Jr.: **Context-dependent control of alternative splicing by RNA-binding proteins.** *Nat Rev Genet* 2014, **15:**689-701.
4. Beyer AL, Osheim YN: **Splice site selection, rate of splicing, and alternative splicing on nascent transcripts.** *Genes Dev* 1988, **2:**754-765.
5. Kornblihtt AR, de la Mata M, Fededa JP, Munoz MJ, Nogues G: **Multiple links between transcription and splicing.** *RNA* 2004, **10:**1489-1498.
6. Carrillo Oesterreich F, Preibisch S, Neugebauer KM: **Global analysis of nascent RNA reveals transcriptional pausing in terminal exons.** *Mol Cell* 2010, **40:**571-581.
7. Khodor YL, Rodriguez J, Abruzzi KC, Tang CH, Marr MT, 2nd, Rosbash M: **Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila.** *Genes Dev* 2011, **25:**2502-2512.
8. Ameur A, Zaghlool A, Halvardson J, Wetterbom A, Gyllensten U, Cavelier L, Feuk L: **Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain.** *Nat Struct Mol Biol* 2011, **18:**1435-1440.
9. Tilgner H, Knowles DG, Johnson R, Davis CA, Chakrabortty S, Djebali S, Curado J, Snyder M, Gingeras TR, Guigo R: **Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs.** *Genome Res* 2012, **22:**1616-1625.
10. Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T: **Regulation of alternative splicing by histone modifications.** *Science* 2010, **327:**996-1000.

11. de la Mata M, Alonso CR, Kadener S, Fededa JP, Blaustein Ma, Pelisch F, Cramer P, Bentley D, Kornblihtt AR: **A Slow RNA Polymerase II Affects Alternative Splicing In Vivo.** *Mol Cell* 2003, **12:**525-532.
12. Roberts GC, Gooding C, Mak HY, Proudfoot NJ, Smith CW: **Co-transcriptional commitment to alternative splice site selection.** *Nucleic Acids Res* 1998, **26:**5568-5572.
13. Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, Oberdoerffer P, Sandberg R, Oberdoerffer S: **CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing.** *Nature* 2011, **479:**74-79.
14. Andersson R, Enroth S, Rada-Iglesias A, Wadelius C, Komorowski J: **Nucleosomes are well positioned in exons and carry characteristic histone modifications.** *Genome Res* 2009, **19:**1732-1741.
15. Hon G, Wang W, Ren B: **Discovery and annotation of functional chromatin signatures in the human genome.** *PLoS Comput Biol* 2009, **5:**e1000566.
16. Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahringer J: **Differential chromatin marking of introns and expressed exons by H3K36me3.** *Nat Genet* 2009, **41:**376-381.
17. Nahkuri S, Taft RJ, Mattick JS: **Nucleosomes are preferentially positioned at exons in somatic and sperm cells.** *Cell Cycle* 2009, **8:**3420-3424.
18. Schwartz S, Meshorer E, Ast G: **Chromatin organization marks exon-intron structure.** *Nat Struct Mol Biol* 2009, **16:**990-995.
19. Spies N, Nielsen CB, Padgett RA, Burge CB: **Biased chromatin signatures around polyadenylation sites and exons.** *Mol Cell* 2009, **36:**245-254.
20. Iannone C, Pohl A, Papasaikas P, Soronellas D, Vicent GP, Beato M, Valcarcel J: **Relationship between nucleosome positioning and progesterone-induced alternative splicing in breast cancer cells.** *RNA* 2015.

21. Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, Valcarcel J, Guigo R: **Nucleosome positioning as a determinant of exon recognition.** *Nat Struct Mol Biol* 2009, **16:**996-1001.
22. Enroth S, Bornelov S, Wadelius C, Komorowski J: **Combinations of histone modifications mark exon inclusion levels.** *PLoS One* 2012, **7:**e29911.
23. **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489:**57-74.
24. **A user's guide to the encyclopedia of DNA elements (ENCODE).** *PLoS Biol* 2011, **9:**e1001046.
25. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al: **Landscape of transcription in human cells.** *Nature* 2012, **489:**101-108.
26. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes.** *Nature* 2008, **456:**470-476.
27. Magen A, Ast G: **The importance of being divisible by three in alternative splicing.** *Nucleic Acids Res* 2005, **33:**5574-5582.
28. Zheng CL, Fu XD, Gribskov M: **Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse.** *RNA* 2005, **11:**1777-1787.
29. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: **High-resolution profiling of histone methylations in the human genome.** *Cell* 2007, **129:**823-837.
30. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS: **Unsupervised pattern discovery in human chromatin structure through genomic segmentation.** *Nat Methods* 2012, **9:**473-476.
31. Kundaje A, Kyriazopoulou-Panagiotopoulou S, Libbrecht M, Smith CL, Raha D, Winters EE, Johnson SM, Snyder M, Batzoglou S, Sidow A: **Ubiquitous heterogeneity and asymmetry of the chromatin**

environment at regulatory elements. *Genome Res* 2012, **22:**1735-1747.

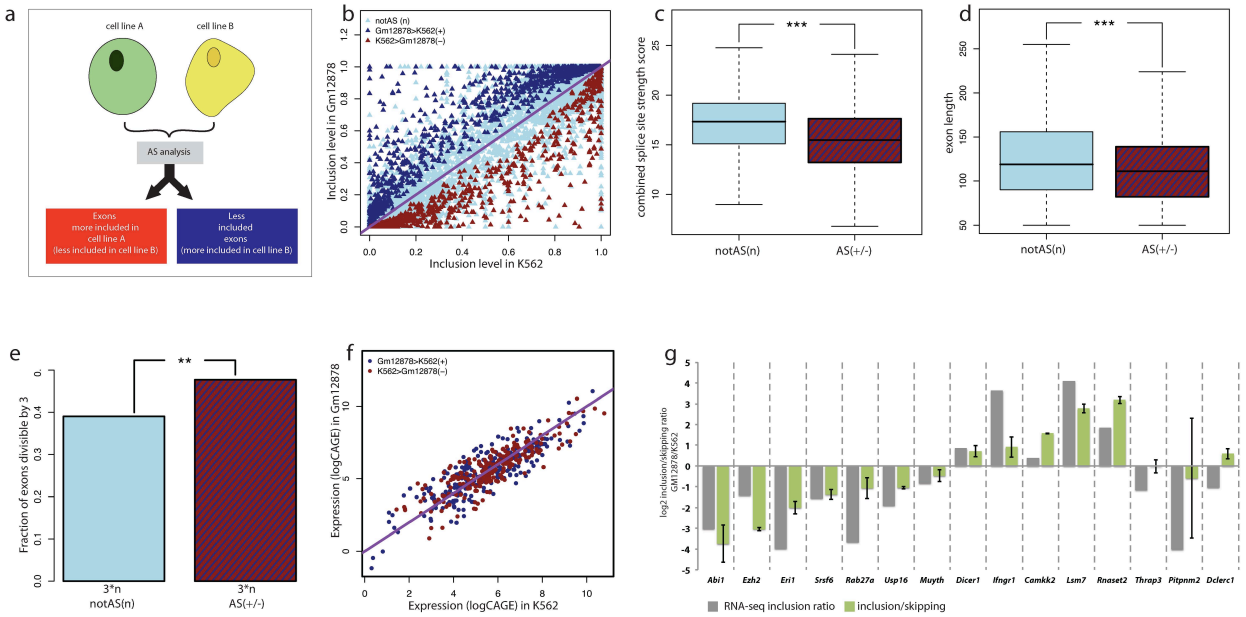
32. Consortium G: **The Genotype-Tissue Expression (GTEx) project.** *Nat Genet* 2013, **45:**580-585.
33. Zhao K, Wang W, Rando OJ, Xue Y, Swiderek K, Kuo A, Crabtree GR: **Rapid and phosphoinositol-dependent binding of the SWI/SNF-like BAF complex to chromatin after T lymphocyte receptor signaling.** *Cell* 1998, **95:**625-636.
34. Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M, et al: **CAGE: cap analysis of gene expression.** *Nat Methods* 2006, **3:**211-222.
35. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, et al: **An oestrogen-receptor-alpha-bound human chromatin interactome.** *Nature* 2009, **462:**58-64.
36. Mercer TR, Edwards SL, Clark MB, Neph SJ, Wang H, Stergachis AB, John S, Sandstrom R, Li G, Sandhu KS, et al: **DNase I-hypersensitive exons colocalize with promoters and distal regulatory elements.** *Nat Genet* 2013, **45:**852-859.
37. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RK, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR, et al: **RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease.** *Science* 2015, **347:**1254806.
38. Lukk M, Kapushesky M, Nikkila J, Parkinson H, Goncalves A, Huber W, Ukkonen E, Brazma A: **A global map of human gene expression.** *Nat Biotechnol* 2010, **28:**322-324.
39. Schor IE, Gomez Acuna LI, Kornblihtt AR: **Coupling between transcription and alternative splicing.** *Cancer Treat Res* 2013, **158:**1-24.
40. de la Mata M, Lafaille C, Kornblihtt AR: **First come, first served revisited: factors affecting the same alternative splicing event have different effects on the relative rates of intron removal.** *RNA* 2010, **16:**904-912.

41. Dujardin G, Lafaille C, de la Mata M, Marasco LE, Munoz MJ, Le Jossic-Corcos C, Corcos L, Kornblihtt AR: **How slow RNA polymerase II elongation favors alternative exon skipping.** *Mol Cell* 2014, **54:**683-690.
42. Kornblihtt AR, Schor IE, Allo M, Dujardin G, Petrillo E, Munoz MJ: **Alternative splicing: a pivotal step between eukaryotic transcription and translation.** *Nat Rev Mol Cell Biol* 2013, **14:**153-165.
43. Allo M, Buggiano V, Fededa JP, Petrillo E, Schor I, de la Mata M, Agirre E, Plass M, Eyras E, Elela SA, et al: **Control of alternative splicing through siRNA-mediated transcriptional gene silencing.** *Nat Struct Mol Biol* 2009, **16:**717-724.
44. Saint-Andre V, Batsche E, Rachez C, Muchardt C: **Histone H3 lysine 9 trimethylation and HP1gamma favor inclusion of alternative exons.** *Nat Struct Mol Biol* 2011, **18:**337-344.
45. Schor IE, Kornblihtt AR: **Playing inside the genes: Intragenic histone acetylation after membrane depolarization of neural cells opens a path for alternative splicing regulation.** *Commun Integr Biol* 2009, **2:**341-343.
46. Dekker J, Marti-Renom MA, Mirny LA: **Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data.** *Nat Rev Genet* 2013, **14:**390-403.
47. de Almeida SF, Grosso AR, Koch F, Fenouil R, Carvalho S, Andrade J, Levezinho H, Gut M, Eick D, Gut I, et al: **Splicing enhances recruitment of methyltransferase HYPB/Setd2 and methylation of histone H3 Lys36.** *Nat Struct Mol Biol* 2011, **18:**977-983.
48. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al: **GENCODE: the reference human genome annotation for The ENCODE Project.** *Genome Res* 2012, **22:**1760-1774.
49. Li JJ, Jiang CR, Brown JB, Huang H, Bickel PJ: **Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform**

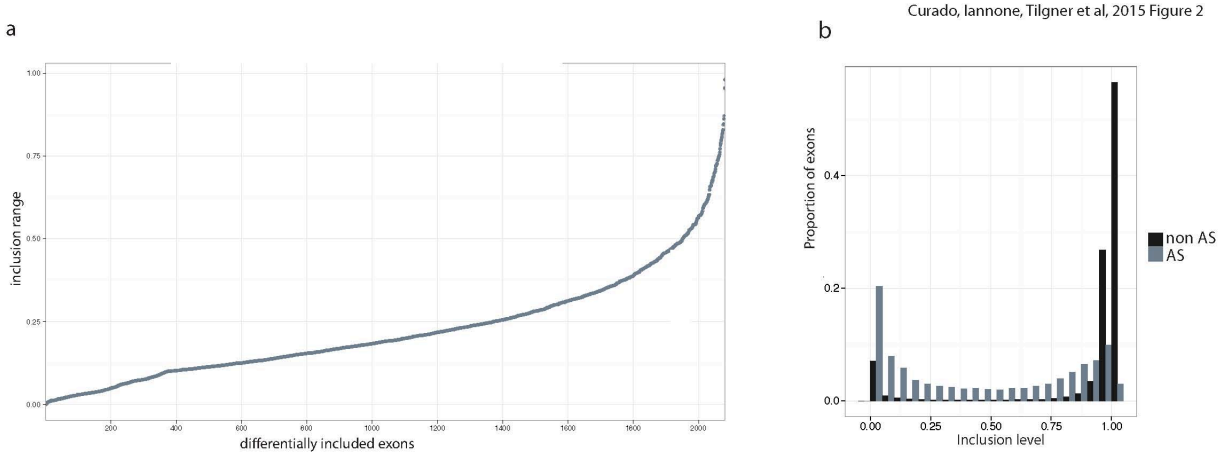
discovery and abundance estimation. *Proc Natl Acad Sci U S A* 2011, **108:**19867-19872.

50. Yeo G, Burge CB: **Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals.** *J Comput Biol* 2004, **11:**377-394.
51. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E, et al: **Integrative annotation of chromatin elements from ENCODE data.** *Nucleic Acids Res* 2013, **41:**827-841.
52. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B: **JASPAR: an open-access database for eukaryotic transcription factor binding profiles.** *Nucleic Acids Res* 2004, **32:**D91-94.
53. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS: **MEME SUITE: tools for motif discovery and searching.** *Nucleic Acids Res* 2009, **37:**W202-208.

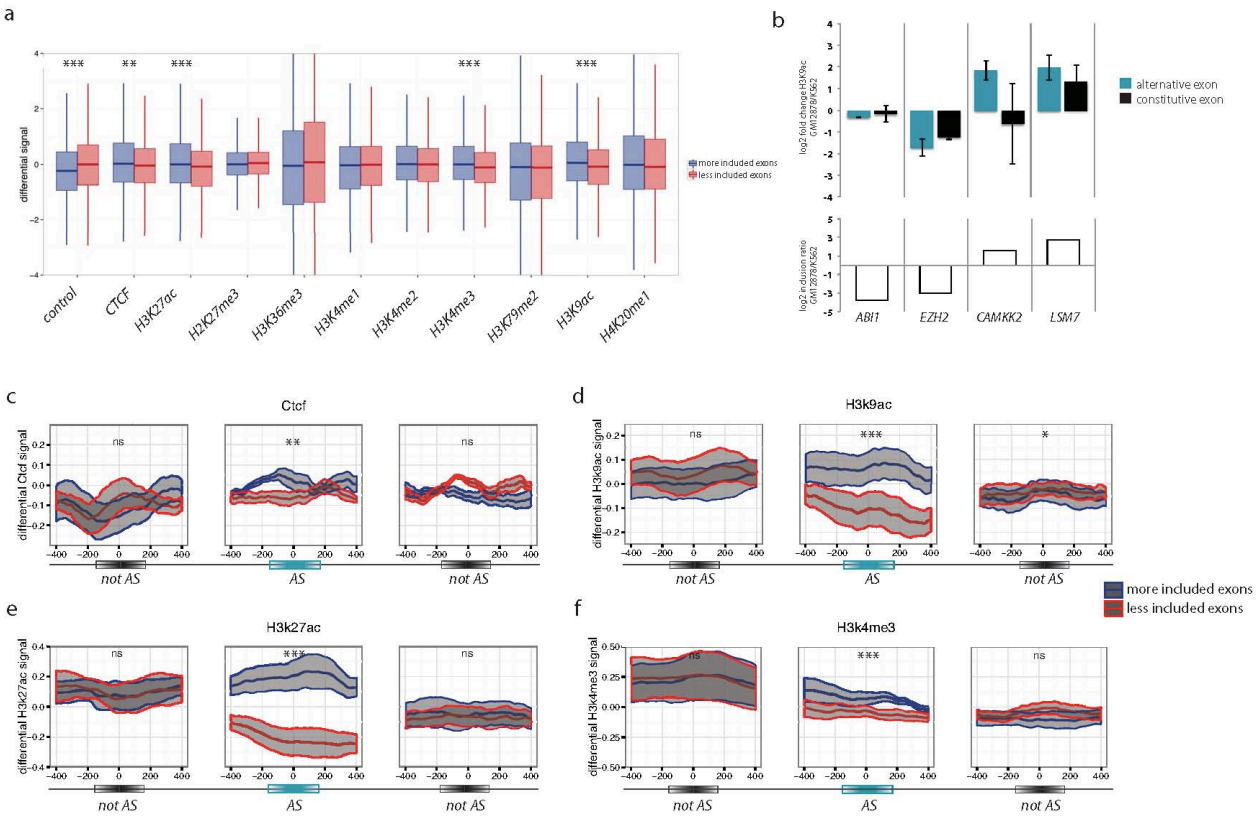
Curado, Iannone, Tilgner et al, 2015 Figure 1



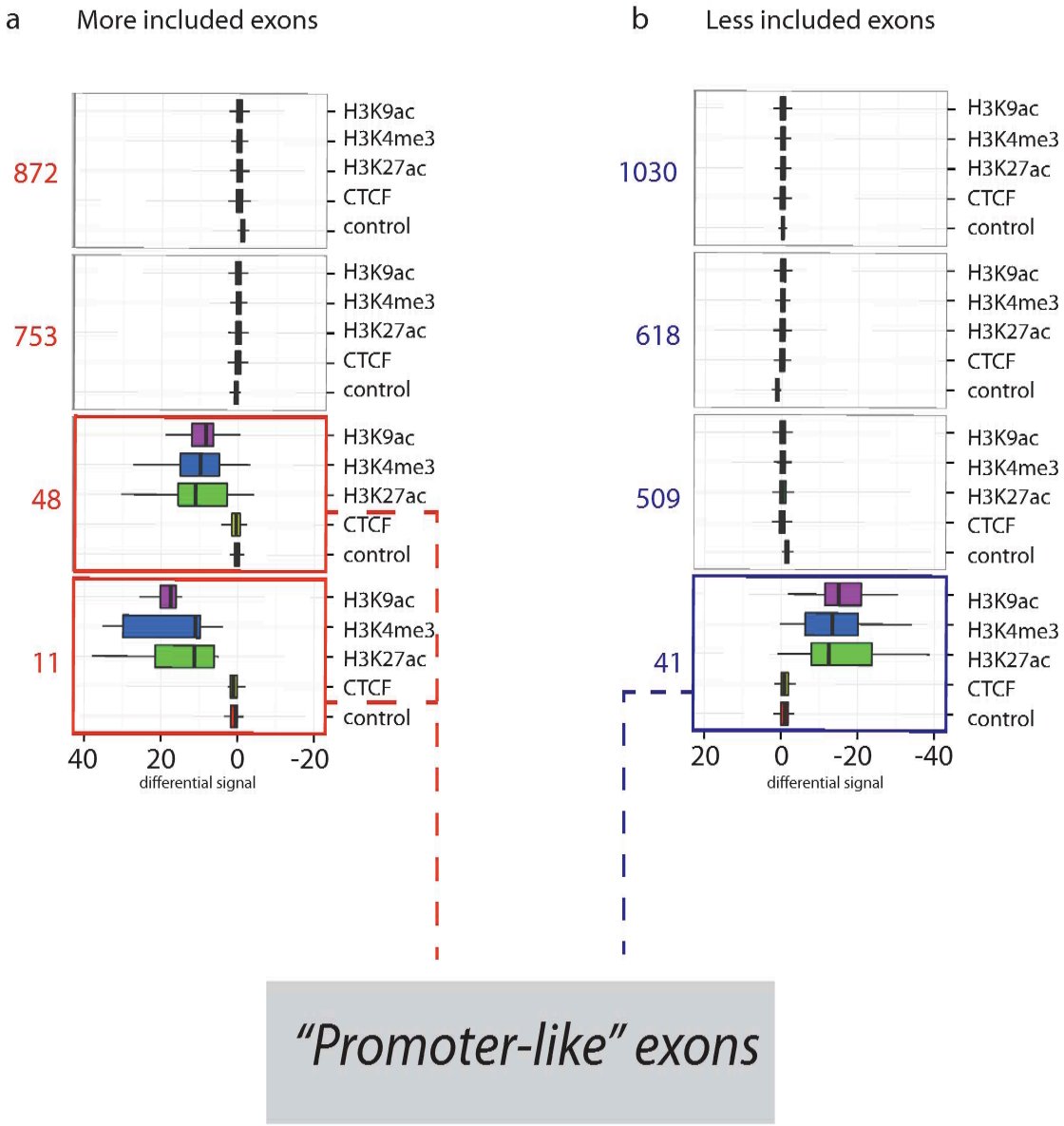
Curado, Iannone, Tilgner et al, 2015 Figure 2



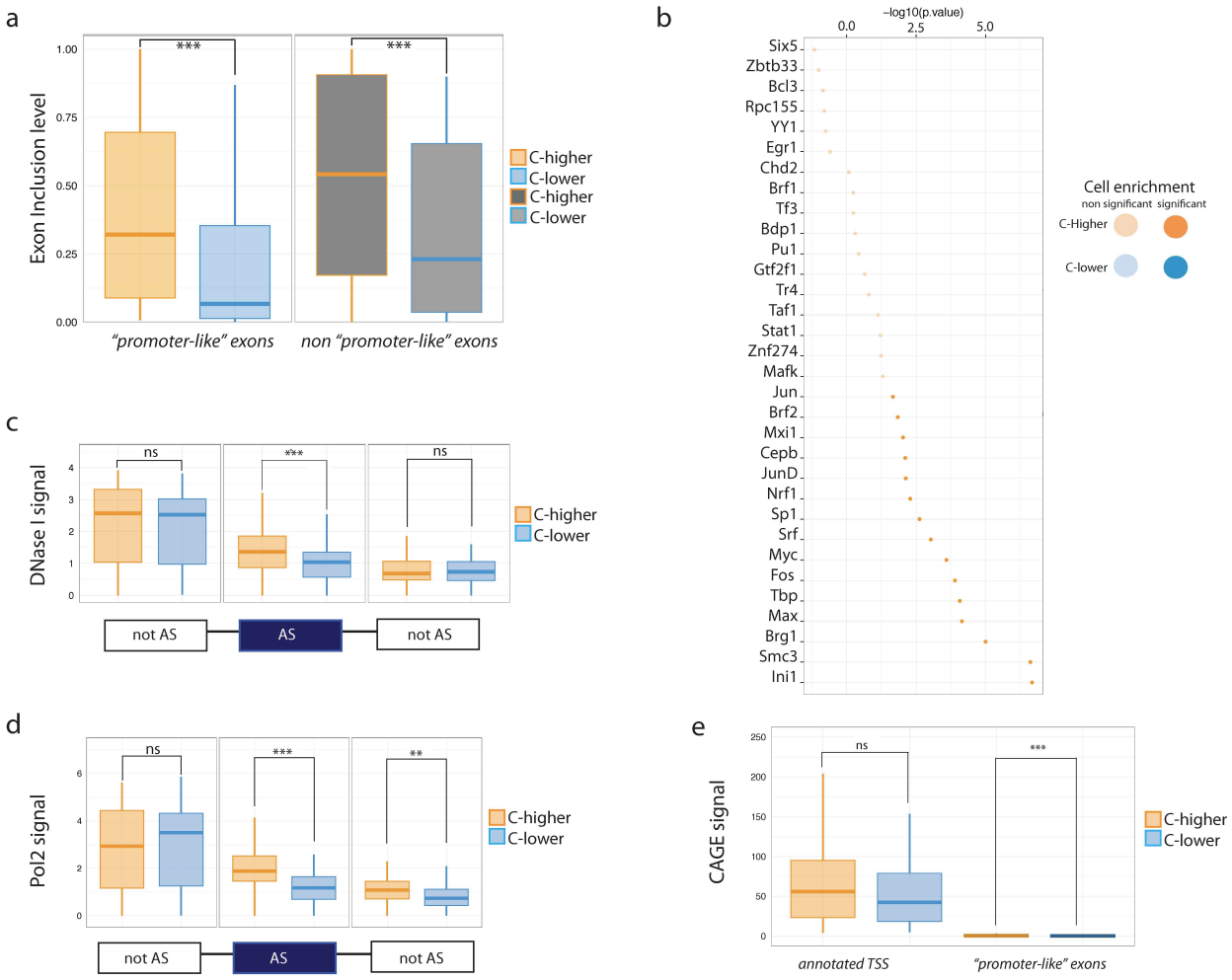
Curado, Iannone, Tilgner et al, 2015 Figure 3



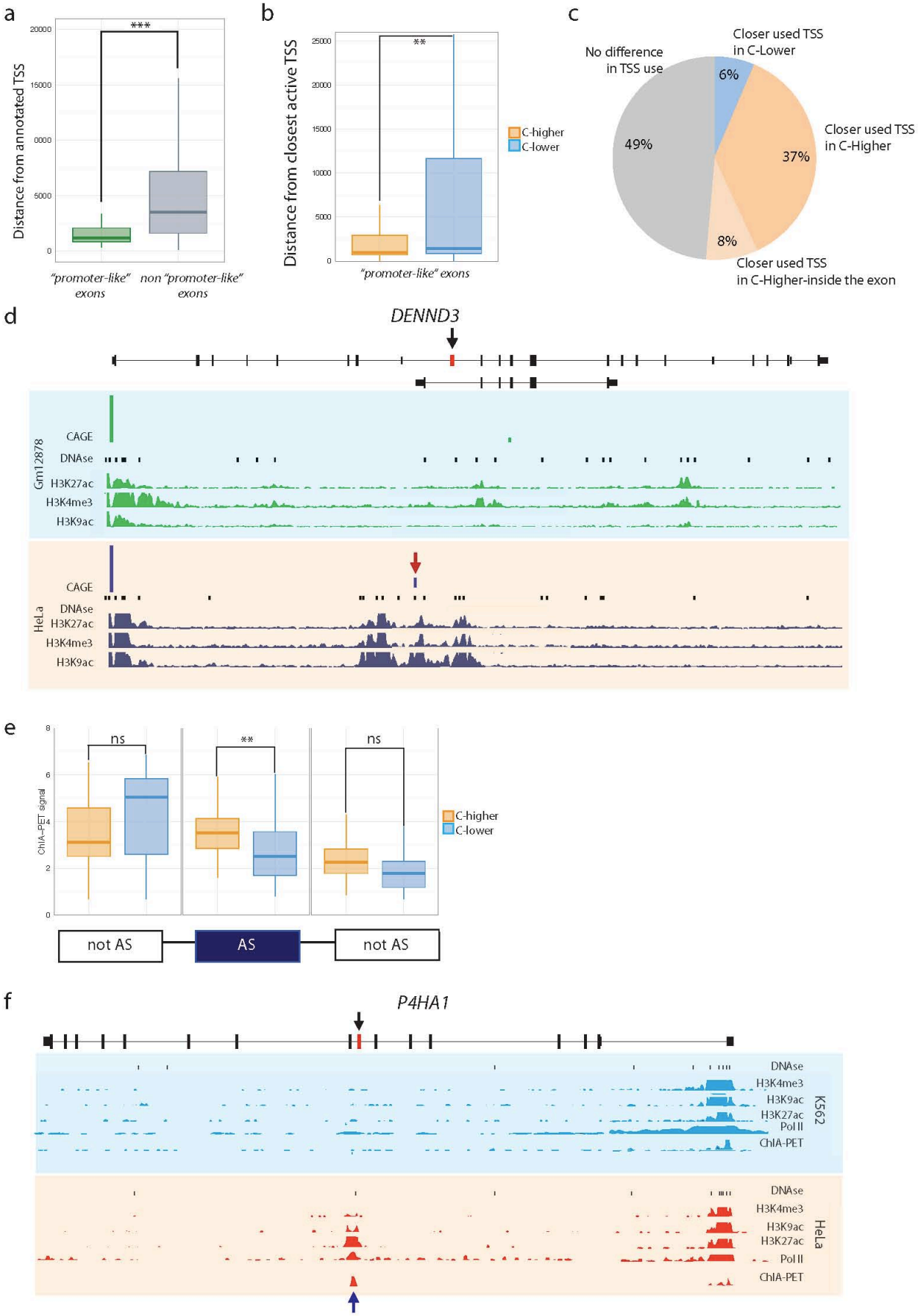
Curado, Iannone, Tilgner et al, 2015 Figure 4



Curado, Iannone, Tilgner et al, 2015 Figure 5



Curado, Iannone, Tilgner et al, 2015 Figure 6



Curado, Iannone, Tilgner et al, 2015 Figure 7

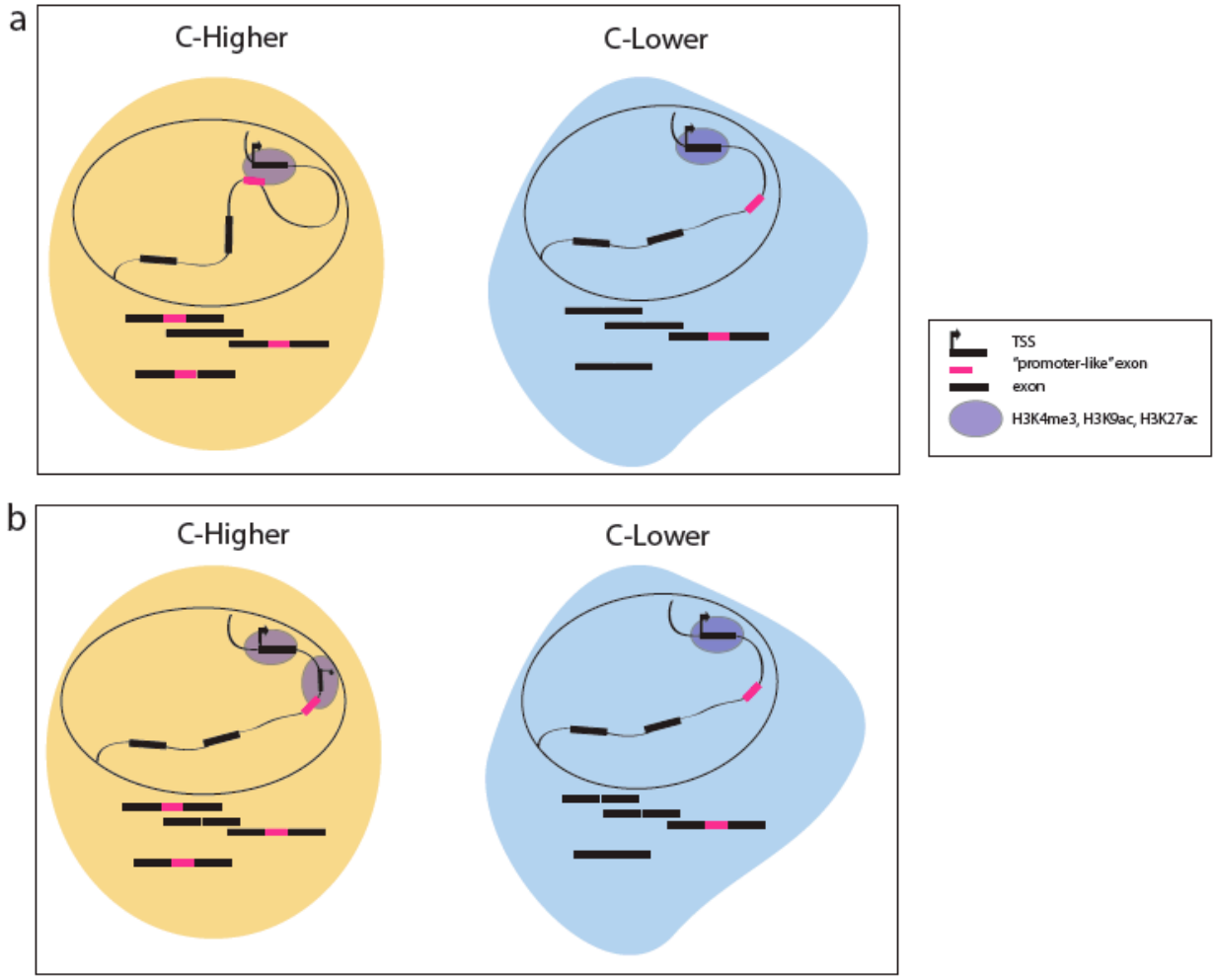


Figure Legends

Figure 1. Assessment and properties of exons differentially spliced between cell lines K562 and Gm12878. (a) Classification of "more included" and "less included" exons. **(b)** Estimated inclusion ratio in the K562 cell line (x-axis) and in the Gm12878 cell line (y-axis) of exons whose inclusion is (i) significantly higher in Gm12878 (dark blue), (ii) significantly higher in the K562 cell line (dark red), (iii) whose inclusion does not change significantly between the two cell lines (light blue). (**c**) Distribution of the sum of the strengths of 5' and 3' splice sites flanking exons that do not display significant differences in inclusion between the cell lines (notAS, left boxplot) and differentially included (i.e. regulated) exons (AS, right boxplot). Wilcoxon rank-sum-tests were calculated for the two distributions, significance levels are indicated: * (0.05>p>0.01), ** (0.01>p>0.001), *** (0.001>p). (**d**) Exon length distribution of AS and notAS exons. (**e**) Fraction of AS and notAS exons, the length of which is multiple of three: only exons that were entirely coding were considered. (**f**) Expression of genes with exons whose inclusion is (i) significantly higher in Gm12878 (dark blue), (ii) significantly higher in the K562 cell line (dark red). X-axis: log2(cage value) in K562; Y-axis: log2(cage value) in Gm12878. **(g)** Experimental validation: comparison between inclusion levels of differentially regulated exons between Gm12878 and K562, calculated analyzing RNASeq ENCODE data (gray bars) or with RT-qPCR analyses of RNA extracted from K562 and Gm12878 (green bars). For RT-qPCR analysis were used primers amplifying specifically the inclusion or skipping isoform. Gray bars represent log2 ratio of inclusion level between Gm12878 and K562 calculated for each exon form RNASeq. Green bars represent log2 ratio between inclusion/skipping ratio in Gm12878 and K562, normalized to the ratio in a constitutive exon on the same transcript; error bars represent standard deviations of three independent experiments. 12 out of 15 exons tested show consistent inclusion direction as measured by RNASeq and RT-qPCR.

Figure 2. Differentially spliced exons in cell lines. (a) Inclusion range of regulated exons. The inclusion range of an exon is defined as the difference between the maximum and the minimum inclusion observed for that exon across the cell lines investigated. Exons are sorted by inclusion range. (**b)** Distribution of the inclusion level of regulated (AS) and non-regulated exons (notAS) across all the cell lines used.

Figure 3. Enrichment of chromatin epigenetic marks on regulated exons. (a) Differential signals (log2, Y-axis) for "more included" (blue) or "less included" (red) exons from the seven cell-pairs used, are represented for 11 ChIPSeq datasets corresponding to different epigenetic marks, CTCF and input DNA (control). The boxplots correspond to the distribution of the average differential signal over the length of each regulated exon. Wilcoxon rank-sum-tests with Bonferroni correction were calculated for the two distributions, significance levels are indicated: * (0.05>p>0.01), ** (0.01>p>0.001), *** (0.001>p). CTCF, H3K9ac, H3K4me3 and H3K27ac have significantly higher signal in "more included" than in "less included" exons, while input DNA control shows the opposite trend. **(b)** Validation of H3K9ac enrichment over H3 by ChIP. Average and standard deviation of log2 of the fold change in H3K9ac signal over total H3 signal in regulated (alternative) and constitutive exons from four different genes. Values are from three independent replicates. White bars in the lower panel represent inclusion level ratios between the two cell lines for the regulated exons, as determined by RNASeq. A general association between exon inclusion and higher levels of H3K9 acetylation is observed. **(c-f)** Differential ChIPSeq signals (average and standard error of the mean) for CTCF, H3K9ac, H3K27ac, and H3K4me3 are represented for "more included" exons (blue) and "less included" exons (red) in a 800bp-window around the middle of the regulated exon (AS) and flanking not regulated (notAS) upstream (left) and downstream (right) exons. Significance levels are indicated by * (0.05>p>0.01), ** (0.01>p>0.001), *** (0.001>p) and ns (p>0.05). Differential accumulation of marks is generally specific of regulated exons.

Figure 4. K-means clustering based on epigenetic signatures of regulated exons. (a, b) boxplots represent differential ChIPSeq signal for more included **(a)** and less included **(b)** exons. Each square represents one group of exons identified by K-means clustering. The number of exons present in each cluster is indicated to the left of each cluster. 3 clusters of exons show strong differential signal for H3K9ac, H3K27ac and H3K4me3 in the direction of the inclusion level change. These exons were called "promoter-like" exons due to the nature of these histone modifications.

Figure 5. Characterization of "promoter-like" exons. (a) Exon inclusion levels in C-higher (yellow) and C-lower (blue) cell lines in "promoter-like" and non "promoter-like" exons. **(b)** Transcription factors binding enrichment significance in C-higher over C-lower cell lines. Bonferroni corrected –log10(p-value) of the enrichment is represented for all the transcription factors tested. All the transcription factors had higher signal in C-higher (orange). Solid colors represents significant enrichments ($p < 0.05$) **(c-d)** DNase I sensitivity and RNA polymerase II signals in promoter-like exons in C-higher and C-lower cell lines. Signals are represented for regulated (AS) and flanking non-regulated (notAS) exons. **(e)** Gene expression levels of exons in C-higher and C-lower conditions measured using CAGE tags. Distributions are given for "bona-fide" annotated TSSs and for "promoter-like" exons location. Significance levels are indicated by * ($0.05>p>0.01$), ** ($0.01>p>0.001$), *** ($0.001>p$) and ns ($p>0.05$).

Figure 6. Relationship between promoter and "promoter-like" exons (a) Distribution of the distance (in nucleotides) between annotated TSS of "promoter-like" and non "promoter-like" exons. **(b)** Distribution of the distance (in nucleotides) between "promoter-like" exons and the nearest active TSS in C-higher and C-lower cell lines. **(c)** Proportion of "promoter-like" exons in which the active TSS is closer in C-higher than in C-lower cell lines, in C-lower than in C-higher cell lines, and at the same distance in C-higher and C-lower cell lines. **(d)** USCS Genome browser view

of the DENND3 gene, that contains a "promoter-like" exon (in red) more included in Hela than in Gm12878 cells. Genomic tracks for CAGE, DNase I and ChIPSeq of H3K9ac, H3K27ac and H3K4me3 levels are displayed. The CAGE signal corresponding to the alternative active promoter, used in Hela, is marked with a red arrow. **(e)** CHIA-PET signal in "promoter-like" exons in C-higher and C-lower cell lines. Signals are represented for regulated (AS) and flanking non-regulated (notAS) exons. Significance levels are indicated by * (0.05>p>0.01), ** (0.01>p>0.001), *** (0.001>p) and ns (p>0.05). (f) USCS Genome browser view of the P4HA1 gene, that contains an exon (in red) more included in Hela than in K562 cells. Genomic tracks for DNase I, ChIA-PET and ChIPSeq of Pol II, H3K9ac, H3K27ac and H3K4me3 are displayed. The ChIA-PET signal, specific of Hela cells, is marked with a blue arrow.

Figure 7. Models linking promoter activity with inclusion of "promoter-like" exons. Looping model (a): physical interactions between TSS and the genomic region corresponding to the alternative exon facilitates exon inclusion (left), while absence of such interactions leads to more skipping of "promoter-like" exons (pink). **Alternative TSS model (b):** the activation of an alternative TSS in the proximity of the genomic region corresponding to the alternative exon facilitates exon inclusion (left), while the inactivation of the TSS closer to the "promoter-like" exon promotes its skipping.

CHAPTER 4 – DISCUSSION

Genomic DNA, in eukaryotes, is packed into nucleosomes, octamers of histone proteins around which DNA is wrapped. Histone tails are subject to a vast array of post-translational modifications including acetylation and methylation. These modifications convert chromatin between “closed” heterochromatin and more “open” euchromatin states creating a constant state of flux. Since most of gene expression steps, from transcription to RNA processing, happen in a chromatin context it is widely accepted that there are connections between them. Levels of histone modifications throughout the gene are associated with regulation at the initiation and elongation steps of transcription (Li, Carey, and Workman) and specific accumulation of certain chromatin marks over the exons were already associated with splicing regulation (Luco and Misteli). However, in most of the cases, evidence remains correlative in nature and we still lack a rigorous mechanistic proof of such event.

In the frame of this thesis work, we aimed to better understand connections between chromatin structure and RNA production. With the advent of high-throughput sequencing techniques great amounts of data was generated, published and shared with the community. Cases like the ENCODE (Dunham et al.) or modENCODE (Roy et al.; Gerstein et al.) consortiums made a great effort in making everything public and ready to be used by the community. In this thesis we took advantage of this opportunity and analyzed both public available and privately generated data in the pursuit of our answers.

In the first section of the results chapter we re-visited gene expression in the context of chromatin by studying the regulation of developmental temporally active genes. We observed that the regulation of these specific genes happens in the absence of canonically active histone modifications, possibly depending more on the binding of transcription factors. At the same time we found that chromatin states correlate with

expression stability (both at the transcription and splicing level) and not only with expression level. Another of our topics of interest is splicing and so, in the second section of the results, we studied the "co-transcriptionality" of RNA splicing. Beyer and Osheim put forward for the first time, in 1988, the idea that splicing could occur co-transcriptionally. Only now we had the tools to confirm that this was actually a very frequent phenomenon. A series of independent, concurrent studies showed that chromatin follows an intron-exon structure and that nucleosomes are especially enriched in exons when compared to exons(Nahkuri, Taft, and Mattick; Schwartz, Meshorer, and Ast; Spies et al.; Tilgner, Nikolaou, et al.). Knowing that splicing, globally, occurs while the pre-mRNA is still tethered to the transcribing RNA polymerase, the role of chromatin in splicing regulation was the logical question that arose right after. In the third section of the results we decided to compare changes in chromatin structure co-occurring with changes in splicing decisions between differentiated cell lines. The goal was to define a set of exons whose splicing decisions were under strong chromatin regulation and to propose a possible model to this process.

A. Transcriptional activation without chromatin marking in developmentally regulated genes

Regulation of gene expression is crucial to maintain cell identity but, throughout the lifetime of an organism, it also needs to be flexible enough to allow for responses to endogenous and exogenous stimuli. Development of an organism is a process that needs quick and accurate regulation and coordination of the cells. Modifications on DNA and on the histone proteins are known to control gene expression by establishing and maintaining specific chromatin states but in recent years it became evident that there were exceptions. Since regulated activation and de-activation of genes is particularly important and precise during morphogenesis, we

decided to study the chromatin context of temporarily active genes and to check if it followed tightly their activation.

In this project we analyzed publicly available modENCODE ChIPSeq and RNASeq data in *D. melanogaster* and *C. elegans* and also generated some additional data for deeper exploration. We reported transcriptional activation of developmental regulated genes in the absence of canonically active chromatin marks. Genes that are temporarily active don't have the typical histone marks of active genes and this finding is strongly supported by targeted validation experiments. These genes seem to have under stronger promoter regulation, as they are under stronger selective constrains, have enrichment of putative sites for transcription factors and also show different transcription factors binding profiles.

Another finding we made was the positive correlation between canonically activator histone marks and RNA producing stability. Genes with stronger active histone modification marking have a more stable expression level, during development. We also reported an association between highly structured chromatin states and regulation of splicing. Indeed, we observed more stochastic production of alternative splicing isoforms in unmarked, developmentally regulated genes than in marked stably expressed ones.

Overall, our results lead us to hypothesize two major transcriptional regulatory programs: In constitutively expressed genes, strong chromatin marking leads to stable transcriptional activity and tightly controlled RNA production with a comparatively smaller role of transcription factors; In genes needing rapidly activation/de-activation a more flexible, unmarked chromatin state is observed with transcription factors assuming the predominant role.

B. Frequency of co-transcriptional splicing in humans

When different steps in mRNA biogenesis occur at the same time and place there are opportunities for coupling or cross talk. During gene expression, if the act of intron removing happens in close proximity and while RNA polymerase is still transcribing the DNA, it opens a wealth of opportunities for chromatin to influence splicing. The possibility for regulation comes not only from the spatial proximity but also because chromatin can influence transcription dynamics, which in turn can influence exon inclusion levels in a feedforward and feedback kind of circuit (Mata et al.; Howe, Kane, and Ares). Here we decided to assess the frequency of co-transcriptional splicing, in the framework of the ENCODE project.

We started by introducing the "completed splicing index" (coSI) for each exon in different RNASeq experiments, coming from different fractions/compartments of the cell. By analyzing RNASeq data coming from the chromatin fraction (the sequenced RNA was still being transcribed and in contact with the chromatin) and the nuclear polyA-fraction (transcribed RNA without polyA tail) we concluded that co-transcriptional splicing is widespread in humans. Splicing also tends to proceed in a 5' to 3' direction, fitting with the "first come first served" rule, although exceptions exist.

Exons showed differences in splice site strength, binding sites for SR proteins and hnRNPs and chromatin organization depending on how much splicing was completed. Since chromatin structure and Pol II occupancy change along the gene body (Barski et al.), as coSI does, the chromatin observation might reflect the position within the gene. This correlation might be functional or completely circumstantial. Nevertheless, decision tree analysis showed that chromatin contains predictive capacity for separating exons with high and low coSI, early and late spliced, respectively. This predictive capacity did not entirely came from the position within the gene (distance to TSS and polyA-site). Completion of splicing in lncRNAs seems to be less efficient than in protein coding genes.

CoSI values of lncRNAs are dramatically lower than those of coding exons in the chromatin fraction and persist in the nuclear polyA+ fraction. LncRNAs are often spliced later or not spliced at all, suggesting a less important role for splicing in non-coding RNAs.

This project presents another piece of evidence that chromatin, transcription and splicing are connected and cannot always be completely separated.

C. Co-occurrence of promoter-like chromatin marks and splicing regulation

Using a statistical framework we carried out pairwise comparisons between differentiated human cell lines. We evaluated changes in splicing and changes in chromatin and tried to identify co-occurrence of both in order to capture exons whose splicing decisions could be influenced by epigenetic features. This link was already studied in the recent past and some histone marks were already reported as connected with splicing by recruiting splicing factors or, indirectly, by modulating RNA polymerase II dynamics in specific exons (Luco, Allo, et al.). With a different approach than these studies, we decided to take advantage of the ENCODE consortium data to study this connection genome-wide. We used polyA+ nuclear RNASeq to identify differentially usage of exons between cell-lines and, to our surprise, only 3% of the evaluated exons could be found as significantly regulated among the five human cell-lines used. These exons are also characterized by a low level of inclusion when compared with the majority of the exons. With ChIPSeq data from different histone modifications we found a positive correlation between accumulation of H3K9ac, H3K27ac, H3K4me3 and CTCF and higher inclusion levels in the regulated exons. We also noticed that the accumulation of this histone marks was local and limited to the alternative exon cassette.

Since we believe that not all internal exons can be susceptible of chromatin regulation during transcription we decided to perform clustering analysis by the chromatin profiles. Strikingly we identified a smaller subset of exons in which histone modification levels changes between cell lines were very dramatic and always in the same direction for all the marks monitored. These exons were highly marked by H3K9ac, H3K27ac and H3K4me3 in the cell line with higher inclusion level and the levels occupancy of these marks was enough to predict splicing changes of these exons even in cell lines not used for the discovery. Since these histone marks are characteristic of promoter regions we called these exons “promoter-like” exons. In this line of thought we also found significant enrichment of RNA polymerase II signal, DNase I hypersensitivity signal and transcription factors binding in the cell-line of higher exon usage.

Interestingly these “promoter-like” exons are also (linear or three dimensional) proximal to the active promoters. Recent work reported that some internal exons are marked by an enhancer-like and promoter-like signature due to a physical association through looping between the exon and its promoter or enhancer (Mercer et al.). We found enrichment of ChIA-PET signals in the cell line of higher inclusion, suggesting the existence of long-range interaction between these exons and the promoters with an influence in splicing regulation.

The accumulation of H3K4me3, H3K27ac and H3K9ac together with stronger DNase hypersensitive sites implies that these exons lie down in regions of open chromatin. Although this evidence should lead to faster elongation rates (and therefore more skipping), the accumulation of RNA polymerase II states for the opposite. We do indeed know that these exons have higher inclusion levels and therefore the stalling of polymerase could be due to some other factors binding in these DHS marked exons. Some of the transcription factors found enriched in conditions of higher inclusion were already implicated in splicing, like the Brg1.

Our results argue for a promoter-associated chromatin role in the regulation of alternative splicing of low usage exons. Mediated by the

proximity to the active transcription start site, either in a linear way or by specific three-dimensional organization of the genome, the local re-arrangements of chromatin in these promoter-like exons can facilitate the binding of external factors, with regulatory functions, that make these exons more included than in the absence of these context. These observations should be further validated and complemented with analysis of the conformation capture technologies to understand the underlying molecular mechanisms.

CHAPTER 5 – CONCLUSIONS

Part I

1. Transcription of temporally active genes during fly and yeast development occurs mostly in the absence of histone modifications canonically linked to gene activation

2. Strong chromatin marking is also associated with expression stability and not only with expression level

3. Highly structured chromatin state favors regulated exon inclusion and, consistently, unmarked genes exhibit a more stochastic alternative splicing pattern than marked stably expressed genes

4. In genes that are constitutively expressed, strong chromatin marking leads to transcriptional stability and tightly controlled RNA production. In these genes, regulation by Transcription Factors would play a comparatively minor role. In contrast, genes that need to be rapidly activated and de-activated are characterized by an unmarked chromatin state that leads to a less regulated, more stochastic RNA production. In these genes, Transcription Factors would play the predominant regulatory role.

Part II

1. Co-transcriptional splicing is a reality for the majority of the human exons

2. Splicing tends to proceed in a 5' to 3' direction, fitting with the "first come first served" rule

3. Most of the splicing events are finished before the polyA tail is added to the transcripts

4. LncRNAs are often spliced later or not spliced at all, suggesting a less important role for splicing in non-coding RNAs.

Part III

1. Only a minority of the internal human exons is differentially spliced between human cell lines. These exons are characterized by a relatively low usage level.

2. From this exon set, only a fraction show a strong correlation between inclusion levels and chromatin levels

3. In these chromatin-associated exons, levels of H3K9ac, H3K27ac, H3K4me3, DNase I hypersensitivity, Pol II and various transcription factors, typical marks of promoters, are found positively associated with their inclusion levels

4. These observations suggest a functional role for transcription-activator chromatin features in the splicing regulation of alternative exons in close proximity

BIBLIOGRAPHY

Agalioti, Theodora, Guoying Chen, and Dimitris Thanos. "Deciphering the Transcriptional Histone Acetylation Code for a Human Gene." *Cell* 111 (2002): 381–392.

Alexander, Ross D. et al. "Splicing-Dependent RNA Polymerase Pausing in Yeast." *Molecular Cell* 40 (2010): 582–593.

Alló, Mariano et al. "Control of Alternative Splicing through siRNA-Mediated Transcriptional Gene Silencing." *Nature structural & molecular biology* 16 (2009): 717–724.

Ameur, Adam et al. "Total RNA Sequencing Reveals Nascent Transcription and Widespread Co-Transcriptional Splicing in the Human Brain." *Nature Structural & Molecular Biology* 18.12 (2011): 1435–1440. Web. 6 Nov. 2011.

Andersson, Robin et al. "Nucleosomes Are Well Positioned in Exons and Carry Characteristic Histone Modifications." *Genome research* 19.10 (2009): 1732–1741. Web. 19 July 2012.

Barash, Yoseph et al. "Deciphering the Splicing Code." *Nature* 465.7294 (2010): 53–9. Web. 1 Mar. 2012.

Barski, Artem et al. "High-Resolution Profiling of Histone Methylations in the Human Genome." *Cell* 129 (2007): 823–837.

Batsché, Eric, Moshe Yaniv, and Christian Muchardt. "The Human SWI/SNF Subunit Brm Is a Regulator of Alternative Splicing." *Nature structural & molecular biology* 13 (2006): 22–29.

Bechtel, Jason M et al. “The Alternative Splicing Mutation Database: A Hub for Investigations of Alternative Splicing Using Mutational Evidence.” *BMC research notes* 1 (2008): 3.

Beckmann, J S, and E N Trifonov. “Splice Junctions Follow a 205-Base Ladder.” *Proceedings of the National Academy of Sciences of the United States of America* 88.6 (1991): 2380–3.

Beyer, A L, and Y N Osheim. “Splice Site Selection, Rate of Splicing, and Alternative Splicing on Nascent Transcripts.” *Genes & development* 2 (1988): 754–765.

Bonnal, Sophie, Luisa Vigevani, and Juan Valcárcel. “The Spliceosome as a Target of Novel Antitumour Drugs.” *Nature reviews. Drug discovery* 11 (2012): 847–59.

Carey, Michael, Bing Li, and Jerry L. Workman. “RSC Exploits Histone Acetylation to Abrogate the Nucleosomal Block to RNA Polymerase II Elongation.” *Molecular Cell* 24 (2006): 481–487.

Carrillo Oesterreich, Fernando, Stephan Preibisch, and Karla M. Neugebauer. “Global Analysis of Nascent Rna Reveals Transcriptional Pausing in Terminal Exons.” *Molecular Cell* 40 (2010): 571–581.

Chen, Mo, and James L Manley. “Mechanisms of Alternative Splicing Regulation: Insights from Molecular and Genomics Approaches.” *Nature reviews. Molecular cell biology* 10.11 (2009): 741–54. Web. 27 May 2014.

Cheng, Chao et al. “A Statistical Framework for Modeling Gene Expression Using Chromatin Features and Application to modENCODE Datasets.” *Genome Biology* 2011: R15.

Churchman, L Stirling, and Jonathan S Weissman. “Nascent Transcript Sequencing Visualizes Transcription at Nucleotide Resolution.” *Nature* 469 (2011): 368–373.

Close, Pierre et al. “DBIRD Complex Integrates Alternative mRNA Splicing with RNA Polymerase II Transcript Elongation.” *Nature* 2012: 386–389.

Corden, J L et al. “A Unique Structure at the Carboxyl Terminus of the Largest Subunit of Eukaryotic RNA Polymerase II.” *Proceedings of the National Academy of Sciences of the United States of America* 82 (1985): 7934–7938.

Cramer, P et al. “Functional Association between Promoter Structure and Transcript Alternative Splicing.” *Proceedings of the National Academy of Sciences of the United States of America* 94 (1997): 11456–11460.

De Almeida, Sérgio F, and Maria Carmo-Fonseca. “Design Principles of Interconnections between Chromatin and Pre-mRNA Splicing.” *Trends in biochemical sciences* 37.6 (2012): 248–53. Web. 10 July 2014.

De Almeida, Sérgio Fernandes et al. “Splicing Enhances Recruitment of Methyltransferase HYPB/Setd2 and Methylation of Histone H3 Lys36.” *Nature structural & molecular biology* 18 (2011): 977–983.

De la Mata, Manuel, Celina Lafaille, and Alberto R Kornblihtt. “First Come, First Served Revisited: Factors Affecting the Same Alternative Splicing Event Have Different Effects on the Relative Rates of Intron Removal.” *RNA (New York, N.Y.)* 16 (2010): 904–912.

Delest, Anna, Tom Sexton, and Giacomo Cavalli. “Polycomb: A Paradigm for Genome Organization from One to Three Dimensions.” *Current Opinion in Cell Biology* 2012: 405–414.

Derrien, Thomas et al. “The GENCODE v7 Catalog of Human Long Noncoding RNAs: Analysis of Their Gene Structure, Evolution, and

Expression." *Genome research* 22.9 (2012): 1775–89. Web. 27 Oct. 2012.

Dhami, Pawandeep et al. "Complex Exon-Intron Marking by Histone Modifications Is Not Determined Solely by Nucleosome Distribution." *PLoS ONE* 5 (2010): n. pag.

Dujardin, Gwendal et al. "How Slow RNA Polymerase II Elongation Favors Alternative Exon Skipping." *Molecular Cell* (2014): 1–8. Web. 1 May 2014.

Dunham, Ian et al. "An Integrated Encyclopedia of DNA Elements in the Human Genome." *Nature* 489.7414 (2012): 57–74. Web. 1 Nov. 2012.

Dye, Michael J., Natalia Gromak, and Nick J. Proudfoot. "Exon Tethering in Transcription by RNA Polymerase II." *Molecular Cell* 21 (2006): 849–859.

Filion, Guillaume J. et al. "Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in Drosophila Cells." *Cell* 143 (2010): 212–224.

Fong, Y W, and Q Zhou. "Stimulatory Effect of Splicing Factors on Transcriptional Elongation." *Nature* 414.6866 (2001): 929–33.

Fox-Walsh, Kristi L, and Klemens J Hertel. "Splice-Site Pairing Is an Intrinsically High Fidelity Process." *Proceedings of the National Academy of Sciences of the United States of America* 106 (2009): 1766–1771.

Gerstein, Mark B et al. "Integrative Analysis of the Caenorhabditis Elegans Genome by the modENCODE Project." *Science (New York, N.Y.)* 330 (2010): 1775–1787.

Green, M R, T Maniatis, and D A Melton. “Human Beta-Globin Pre-mRNA Synthesized in Vitro Is Accurately Spliced in Xenopus Oocyte Nuclei.” *Cell* 32 (1983): 681–694.

Gunderson, Felizza Q, and Tracy L Johnson. “Acetylation by the Transcriptional Coactivator Gcn5 Plays a Novel Role in Co-Transcriptional Spliceosome Assembly.” *PLoS genetics* 5.10 (2009): e1000682. Web. 15 Nov. 2012.

Hodges, Emily et al. “High Definition Profiling of Mammalian DNA Methylation by Array Capture and Single Molecule Bisulfite Sequencing.” *Genome research* 19 (2009): 1593–1605.

Hödl, Martina, and Konrad Basler. “Transcription in the Absence of Histone H3.2 and H3K4 Methylation.” *Current Biology* (2012): 2253–2257. Web. 11 Nov. 2012.

Howe, Kenneth James, Caroline M Kane, and Manuel Ares. “Perturbation of Transcription Elongation Influences the Fidelity of Internal Exon Inclusion in Saccharomyces Cerevisiae.” *RNA (New York, N.Y.)* 9 (2003): 993–1006.

Kadener, Sebastian et al. “Antagonistic Effects of T-Ag and VP16 Reveal a Role for RNA Pol II Elongation on Alternative Splicing.” *EMBO Journal* 20 (2001): 5759–5768.

Kadener, Sebastián et al. “Regulation of Alternative Splicing by a Transcriptional Enhancer through RNA Pol II Elongation.” *Proceedings of the National Academy of Sciences of the United States of America* 99 (2002): 8185–8190.

Kelemen, Olga et al. “Function of Alternative Splicing.” *Gene* 514.1 (2013): 1–30. Web. 5 June 2014.

Keren-Shaul, Hadas, Galit Lev-Maor, and Gil Ast. “Pre-mRNA Splicing Is a Determinant of Nucleosome Organization.” *PLoS ONE* 8.1 (2013): e53506. Web. 14 Feb. 2013.

Khodor, Yevgenia L., Jerome S. Menet, et al. “Cotranscriptional Splicing Efficiency Differs Dramatically between Drosophila and Mouse.” *RNA (New York, N.Y.)* (2012): 2174–2186. Web. 29 Oct. 2012.

Khodor, Yevgenia L., Joseph Rodriguez, et al. “Nascent-Seq Indicates Widespread Cotranscriptional Pre-mRNA Splicing in Drosophila.” *Genes & Development* 25.23 (2011): 2502–2512. Web. 8 Dec. 2011.

Kim, Eddo, Alon Magen, and Gil Ast. “Different Levels of Alternative Splicing among Eukaryotes.” *Nucleic Acids Research* 35 (2007): 125–131.

Kim, Soojin et al. “Pre-mRNA Splicing Is a Determinant of Histone H3K36 Methylation.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (2011): 13564–13569.

Kireeva, Maria L. et al. “Nature of the Nucleosomal Barrier to RNA Polymerase II.” *Molecular Cell* 18 (2005): 97–108.

Klinck, Roscoe et al. “Multiple Alternative Splicing Markers for Ovarian Cancer.” *Cancer research* 68 (2008): 657–663.

Kolasinska-Zwierz, Paulina et al. “Differential Chromatin Marking of Introns and Expressed Exons by H3K36me3.” *Nature genetics* 41.3 (2009): 376–381. Web. 23 July 2011.

Kornblihtt, Alberto R et al. “Alternative Splicing: A Pivotal Step between Eukaryotic Transcription and Translation.” *Nature reviews. Molecular cell biology* 14.3 (2013): 153–65. Web. 27 May 2014.

Kornblihtt, Alberto R. “Promoter Usage and Alternative Splicing.” *Current Opinion in Cell Biology* 2005: 262–268.

Kouzarides, Tony. "Chromatin Modifications and Their Function." *Cell* 128.4 (2007): 693–705. Web. 23 May 2014.

Krawczak, Michael, Jochen Reiss, and David N. Cooper. "The Mutational Spectrum of Single Base-Pair Substitutions in mRNA Splice Junctions of Human Genes: Causes and Consequences." *Human Genetics* 90 (1992): 41–54.

Kwak, Hojoong, and John T Lis. "Control of Transcriptional Elongation." *Annual review of genetics* 47 (2013): 483–508.

Li, Bing, Michael Carey, and Jerry L. Workman. "The Role of Chromatin during Transcription." *Cell* 128.4 (2007): 707–19. Web. 23 May 2014.

Listerman, Imke, Aparna K Sapra, and Karla M Neugebauer. "Cotranscriptional Coupling of Splicing Factor Recruitment and Precursor Messenger RNA Splicing in Mammalian Cells." *Nature structural & molecular biology* 13 (2006): 815–822.

Liu, H X et al. "A Mechanism for Exon Skipping Caused by Nonsense or Missense Mutations in BRCA1 and Other Genes." *Nature genetics* 27 (2001): 55–58.

Loomis, Rebecca J. et al. "Chromatin Binding of SRp20 and ASF/SF2 and Dissociation from Mitotic Chromosomes Is Modulated by Histone H3 Serine 10 Phosphorylation." *Molecular Cell* 33 (2009): 450–461.

Luco, Reini F, Mariano Allo, et al. "Epigenetics in Alternative Pre-mRNA Splicing." *Cell* 144.1 (2011): 16–26. Web. 17 July 2011.

Luco, Reini F, Qun Pan, et al. "Regulation of Alternative Splicing by Histone Modifications." *Science (New York, N.Y.)* 327.5968 (2010): 996–1000. Web. 7 July 2011.

Luco, Reini F, and Tom Misteli. "More than a Splicing Code: Integrating the Role of RNA, Chromatin and Non-Coding RNA in Alternative Splicing

Regulation." *Current opinion in genetics & development* 21.4 (2011): 366–72. Web. 2 Mar. 2012.

Martins, Sandra Bento et al. "Spliceosome Assembly Is Coupled to RNA Polymerase II Dynamics at the 3′ End of Human Genes." *Nature Structural & Molecular Biology* 2011: 1115–1123.

Mata, Manuel De et al. "A Slow RNA Polymerase II Affects Alternative Splicing In Vivo." *Molecular Cell* 12 (2003): 525–532.

McCracken, S et al. "The C-Terminal Domain of RNA Polymerase II Couples mRNA Processing to Transcription." *Nature* 385 (1997): 357–361.

Mercer, Tim R et al. "DNase I-Hypersensitive Exons Colocalize with Promoters and Distal Regulatory Elements." *Nature genetics* 45.8 (2013): 852–9. Web. 13 Dec. 2013.

Modrek, Barmak, and Christopher Lee. "A Genomic View of Alternative Splicing." *Nature genetics* 30.1 (2002): 13–9.

Müller, Ferenc, and Làszlò Tora. "Chromatin and DNA Sequences in Defining Promoters for Transcription Initiation." *Biochimica et biophysica acta* 1839.3 (2014): 118–28. Web. 23 May 2014.

Muñoz, Manuel J. et al. "DNA Damage Regulates Alternative Splicing through Inhibition of RNA Polymerase II Elongation." *Cell* 137 (2009): 708–720.

Nahkuri, Satu, Ryan J Taft, and John S Mattick. "Nucleosomes Are Preferentially Positioned at Exons in Somatic and Sperm Cells." *Cell cycle (Georgetown, Tex.)* 8 (2009): 3420–3424.

Nechaev, Sergei, and Karen Adelman. "Pol II Waiting in the Starting Gates: Regulating the Transition from Transcription Initiation into Productive Elongation." *Biochimica et biophysica acta* 1809.1 (2011): 34–45. Web. 5 June 2014.

Nègre, Nicolas et al. “A Cis-Regulatory Map of the Drosophila Genome.” *Nature* 471 (2011): 527–531.

Nogues, Guadalupe et al. “Transcriptional Activators Differ in Their Abilities to Control Alternative Splicing.” *The Journal of biological chemistry* 277 (2002): 43110–43114.

Pan, Qun et al. “Deep Surveying of Alternative Splicing Complexity in the Human Transcriptome by High-Throughput Sequencing.” *Nature genetics* 40.12 (2008): 1413–5. Web. 25 May 2014.

Pandya-Jones, Amy, and Douglas L Black. “Co-Transcriptional Splicing of Constitutive and Alternative Exons.” *RNA (New York, N.Y.)* 15 (2009): 1896–1908.

Piacentini, Lucia et al. “Heterochromatin Protein 1 (HP1a) Positively Regulates Euchromatic Gene Expression through RNA Transcript Association and Interaction with hnRNPs in Drosophila.” *PLoS Genetics* 5 (2009): n. pag.

Podlaha, Ondrej et al. “Histone Modifications Are Associated with Transcript Isoform Diversity in Normal and Cancer Cells.” Ed. Alexandre V. Morozov. *PLoS Computational Biology* 10.6 (2014): e1003611. Web. 5 June 2014.

Ries, David, and Michael Meisterernst. “Control of Gene Transcription by Mediator in Chromatin.” *Seminars in Cell and Developmental Biology* 2011: 735–740.

Roberts, G C et al. “Co-Transcriptional Commitment to Alternative Splice Site Selection.” *Nucleic acids research* 26 (1998): 5568–5572. Print.

Romero, Pedro R et al. “Alternative Splicing in Concert with Protein Intrinsic Disorder Enables Increased Functional Diversity in Multicellular

Organisms." *Proceedings of the National Academy of Sciences of the United States of America* 103 (2006): 8390–8395.

Roy, Sushmita et al. "Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE." *Science (New York, N.Y.)* 330.6012 (2010): 1787–97. Web. 10 June 2011.

Schor, Ignacio E et al. "Neuronal Cell Depolarization Induces Intragenic Chromatin Modifications Affecting NCAM Alternative Splicing." *Proceedings of the National Academy of Sciences of the United States of America* 106 (2009): 4325–4330.

Schwartz, Schraga, Eran Meshorer, and Gil Ast. "Chromatin Organization Marks Exon-Intron Structure." *Nature structural & molecular biology* 16.9 (2009): 990–995. Web. 15 July 2011.

Shieh, Grace S et al. "H2B Ubiquitylation Is Part of Chromatin Architecture That Marks Exon-Intron Structure in Budding Yeast." *BMC Genomics* 2011: 627.

Shukla, Sanjeev et al. "CTCF-Promoted RNA Polymerase II Pausing Links DNA Methylation to Splicing." *Nature* (2011): n. pag. Web. 2 Oct. 2011.

Sims, Robert J. et al. "Recognition of Trimethylated Histone H3 Lysine 4 Facilitates the Recruitment of Transcription Postinitiation Factors and Pre-mRNA Splicing." *Molecular Cell* 28 (2007): 665–676.

Smallwood, A. et al. "CBX3 Regulates Efficient RNA Processing Genome-Wide." *Genome Research* 2012: 1426–1436.

Smolle, Michaela, and Jerry L Workman. "Transcription-Associated Histone Modifications and Cryptic Transcription." *Biochimica et biophysica acta* 1829.1 (2013): 84–97. Web. 30 May 2014.

Spies, Noah et al. “Biased Chromatin Signatures around Polyadenylation Sites and Exons.” *Molecular Cell* 36.2 (2009): 245–254. Web. 22 July 2011.

Tilgner, Hagen, David G. Knowles, et al. “Deep Sequencing of Subcellular RNA Fractions Shows Splicing to Be Predominantly Co-Transcriptional in the Human Genome but Inefficient for lncRNAs.” *Genome research* 22.9 (2012): 1616–25. Web. 27 Oct. 2012.

Tilgner, Hagen, Christoforos Nikolaou, et al. “Nucleosome Positioning as a Determinant of Exon Recognition.” *Nature structural & molecular biology* 16.9 (2009): 996–1001. Web. 21 July 2011.

Vargas, Diana Y. Y. et al. “Single-Molecule Imaging of Transcriptionally Coupled and Uncoupled Splicing.” *Cell* 147.5 (2011): 1054–1065. Web. 23 Nov. 2011.

Venables, Julian P et al. “Cancer-Associated Regulation of Alternative Splicing.” *Nature structural & molecular biology* 16 (2009): 670–676.

Wahl, Markus C, Cindy L Will, and Reinhard Lührmann. “The Spliceosome: Design Principles of a Dynamic RNP Machine.” *Cell* 136.4 (2009): 701–18. Web. 26 May 2014.

Wang, Peng et al. “Structural Genomics Analysis of Alternative Splicing and Application to Isoform Structure Modeling.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (2005): 18920–18925.

Wang, Zefeng, and Christopher B Burge. “Splicing Regulation : From a Parts List of Regulatory Elements to an Integrated Splicing Code Splicing Regulation : From a Parts List of Regulatory Elements to an Integrated Splicing Code.” *RNA* 14 (2008): 802–813.

Zhang, Chaolin, Adrian R. Krainer, and Michael Q. Zhang. "Evolutionary Impact of Limited Splicing Fidelity in Mammalian Genes." *Trends in Genetics* 2007: 484–488.

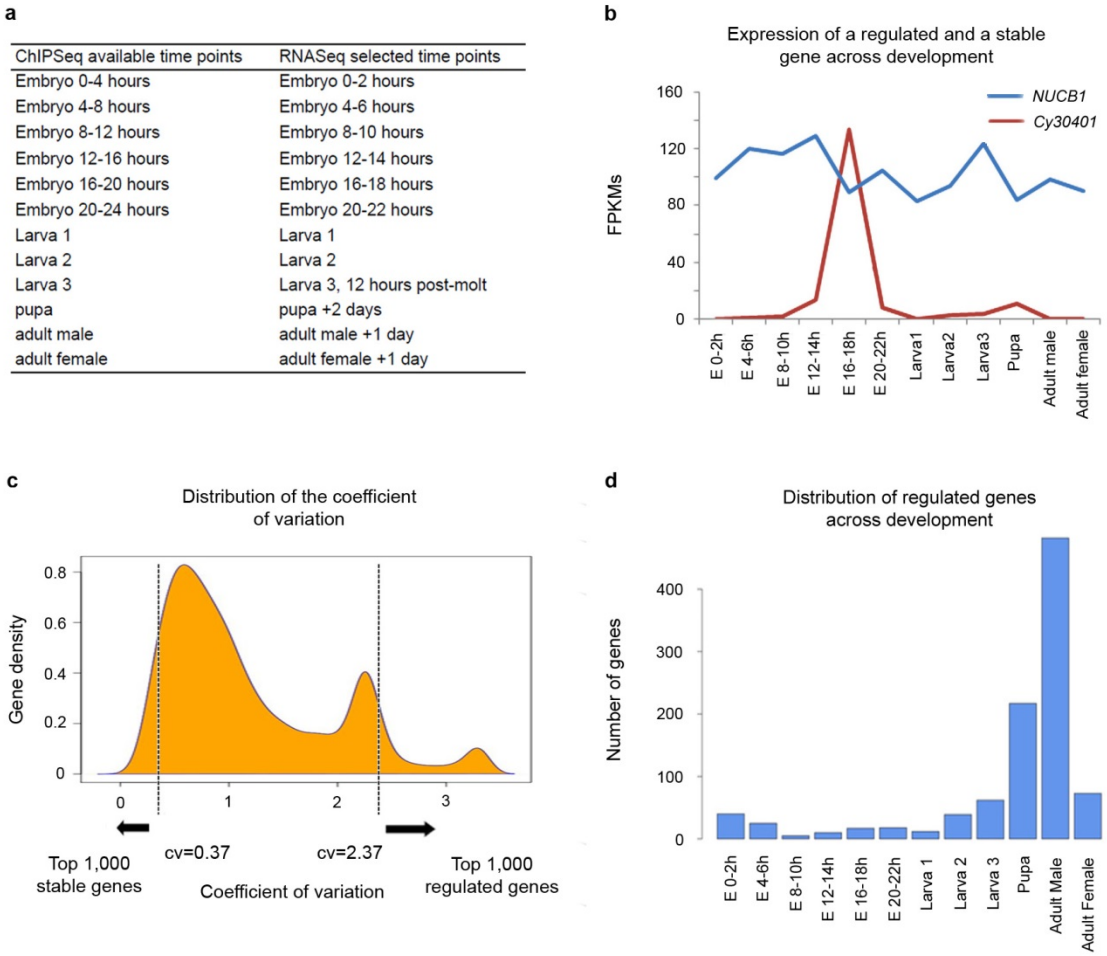
Zhou, Hua-Lin et al. "Hu Proteins Regulate Alternative Splicing by Inducing Localized Histone Hyperacetylation in an RNA-Dependent Manner." *Proceedings of the National Academy of Sciences of the United States of America* 108.36 (2011): E627–35. Web. 31 May 2013.

SUPPLEMENTARY MATHERIAL

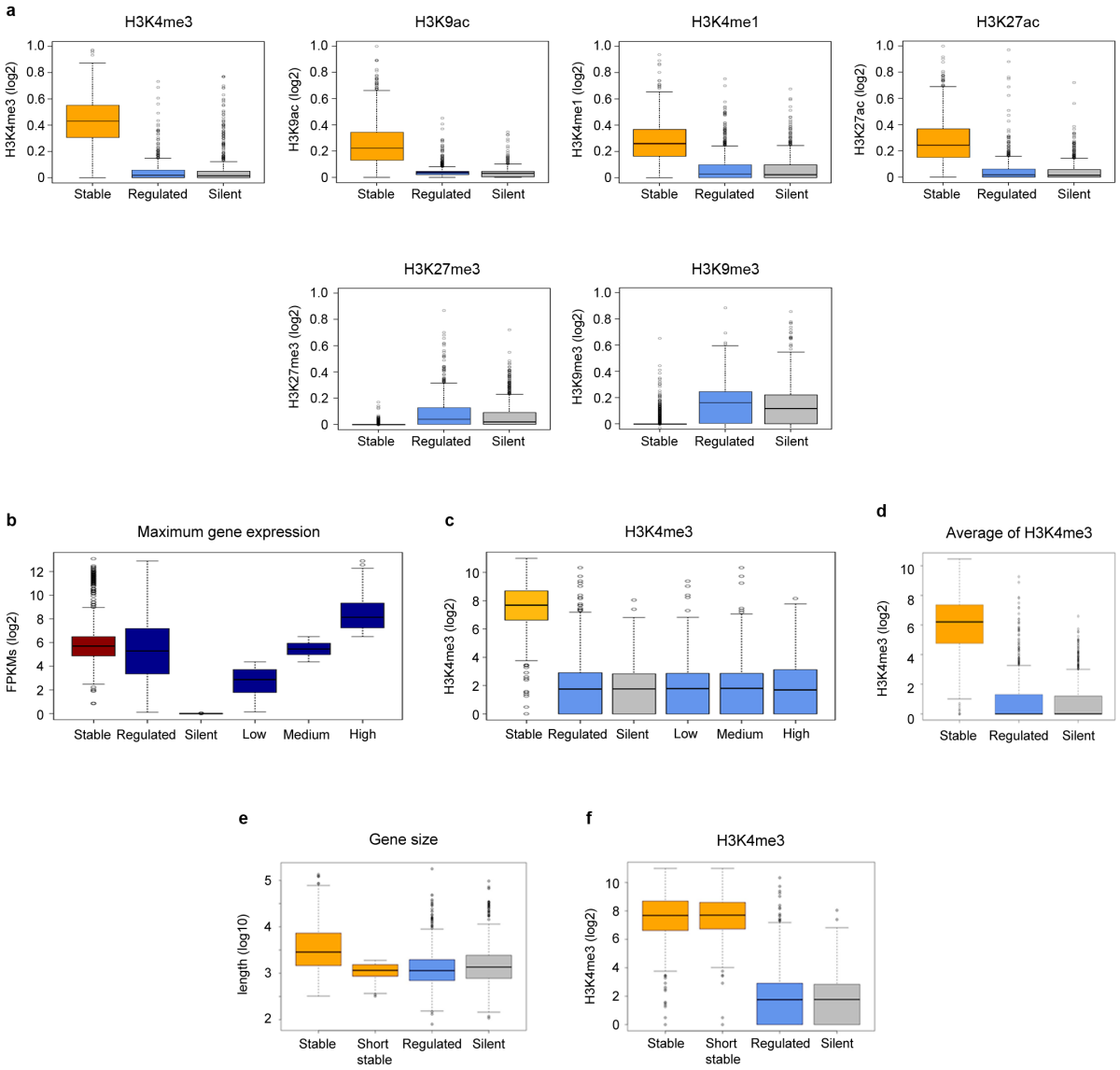
Gene expression without canonical chromatin marking in developmentally regulated genes

a

ChIPSeq available time points	RNASeq selected time points
Embryo 0-4 hours	Embryo 0-2 hours
Embryo 4-8 hours	Embryo 4-6 hours
Embryo 8-12 hours	Embryo 8-10 hours
Embryo 12-16 hours	Embryo 12-14 hours
Embryo 16-20 hours	Embryo 16-18 hours
Embryo 20-24 hours	Embryo 20-22 hours
Larva 1	Larva 1
Larva 2	Larva 2
Larva 3	Larva 3, 12 hours post-molt
pupa	pupa +2 days
adult male	adult male +1 day
adult female	adult female +1 day

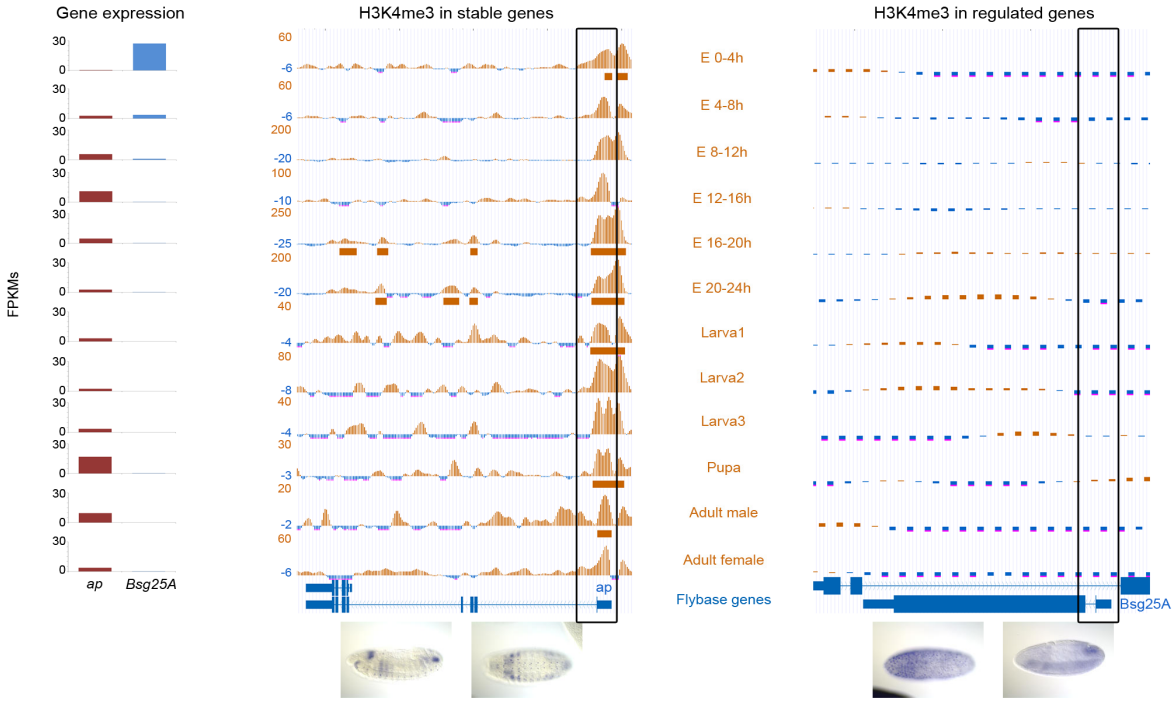


Supplementary Figure 1: Developmentally stable and regulated genes in *D. melanogaster*. **a**, Time points selected for the analysis of chromatin marking in genes regulated during fly development. From the available modENCODE RNASeq data, we selected the 12 points for which ChIPSeq experiments on histone modifications were also available.. **b**, Expression of one stable (*NUCB1)* and one developmentally regulated gene *(Cy30401)*. The value of the coefficient of variation (cv) for *NUCB1* is 0.15. *Cy30401* in contrast, shows a peak of expression in one embryonic stage and, consistently, its cv is 2.49. **c**, Distribution of the coefficient of variation on fly genes. We have calculated the cv of expression for the 12,867 genes for which modENCODE has expression data along *Drosophila* development. The cv distribution uncovers a large class of genes with low coefficient of variation (constant expression during development), and two other minor classes containing genes whose expression is highly variable during development—often restricted to a limited set of stages. For most of the analysis we arbitrarily considered the top 1,000 genes with the lowest cv as stable, and, the top 1,000 genes with the highest cv as developmentally regulated genes (Supplementary Fig. 4). **d**, Time point of maximum expression of developmentally regulated genes.

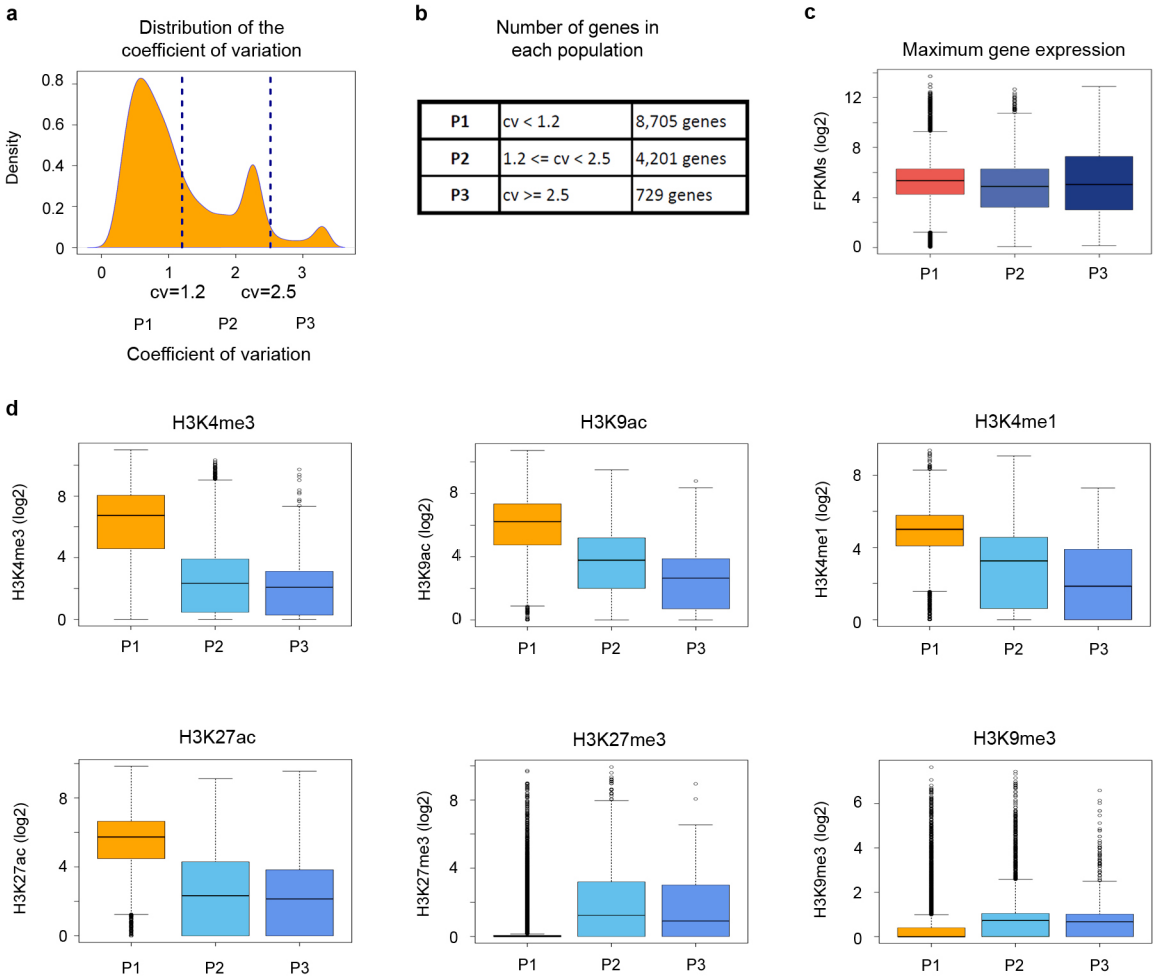


Supplementary Figure 2: Chromatin marking at stable, regulated and silent genes. We performed a number of controls to rule out that our observations arise from undetected confounding factors **a,** Normalized levels of H3K4me3, H3K9ac, H3K4me1, H3K27ac, H3K27me3 and H3K9me3 at the time point of maximum expression during *D. melanogaster* development. Since there are differences in the heights between modENCODE ChIPSeq tracks, we first identified the highest peak of each mark in the genome by checking all the expressed genes and next we used this value to normalize the corresponding profiles. The distributions correspond to the maximum height of the ChIPSeq peak within the gene body for H3K4me3, H3K9ac, H3K4me1 and H3K27ac, and the average height of the ChIPSeq signal over the gene body for H3K27me3 and H3K9me3. Patterns are the same, or even stronger, than those in Figure 1b. **b,** Distribution of expression of top 1,000 stable, top 1,000 regulated, 1,000 silent genes, and of the set of top 1,000 regulated genes divided into three groups according to expression (low, medium, high) at the time point of maximum expression for each gene during fly development. Gene expression was computed as FPKMs by the modENCODE consortium. This was done to control for the differences in the dispersion of gene expression between stable and regulated genes. **c,** Levels of H3K4me3 at the time point of maximum expression during development.for the gene sets defined in

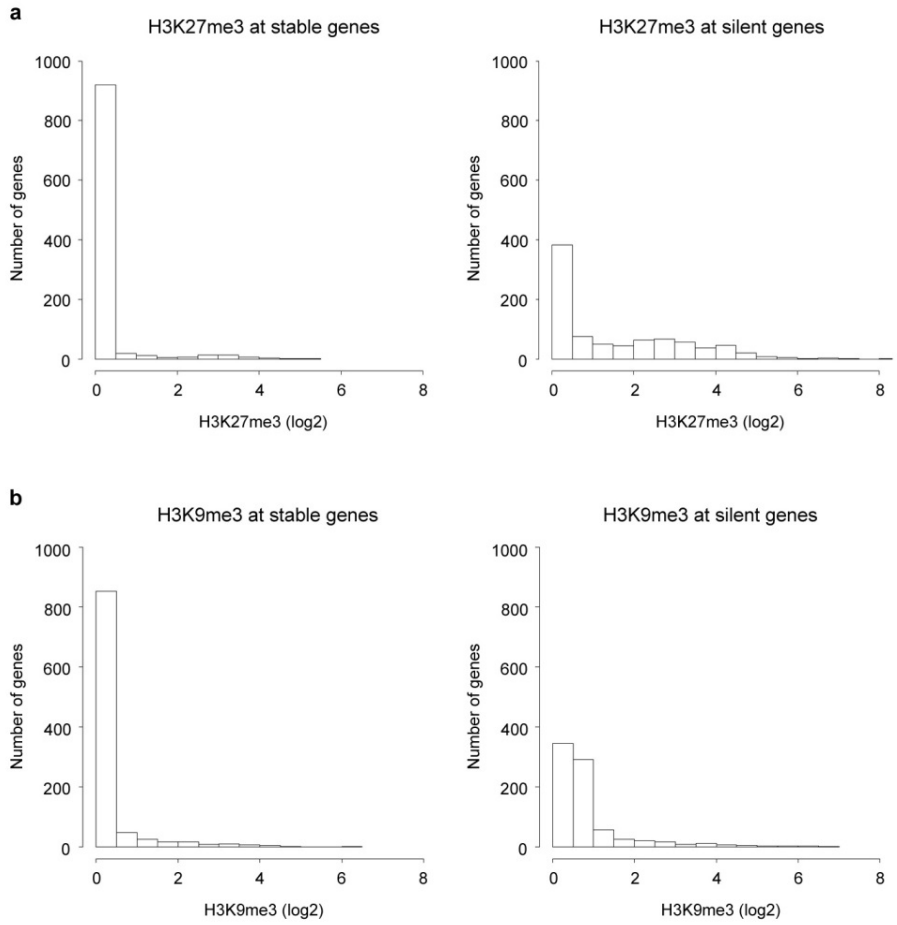
b. Values represent the maximum height of the ChIPSeq peak within the gene body. There is no effect of differences in dispersion of expression, or of the magnitud of expression itself in the patterns observed in Figure 1b **d,** Levels of H3K4me3 at the time point of maximum gene expression computed as the average signal over the gene body, instead of as the maximum peak. The pattern is the same than that in Figure 1b. **e**, Length of stable and regulated genes. Developmentally regulated genes have less number of exons than stable genes (2.6 vs 5.7 on average) and shorter introns (600 bp vs 1,000 bp) and, as a consequence, regulated genes are shorter than stable genes (1,136 bp vs 2,864 bp). To rule out that gene size is a confounding factor, we selected the 520 shortest stable genes. These have an average length (1,188 bp) and number of exons (2.6) very similar to that of variable genes. The figure shows the size distribution of genes in the different classes. **f**, H3K4me3 maximum peak at short stable genes is comparable to the peak at the previous set of stable genes, and it is much higher than the H3K4me3 peak at regulated genes. Therefore, there is no effect of gene length and number of exons in our observations.



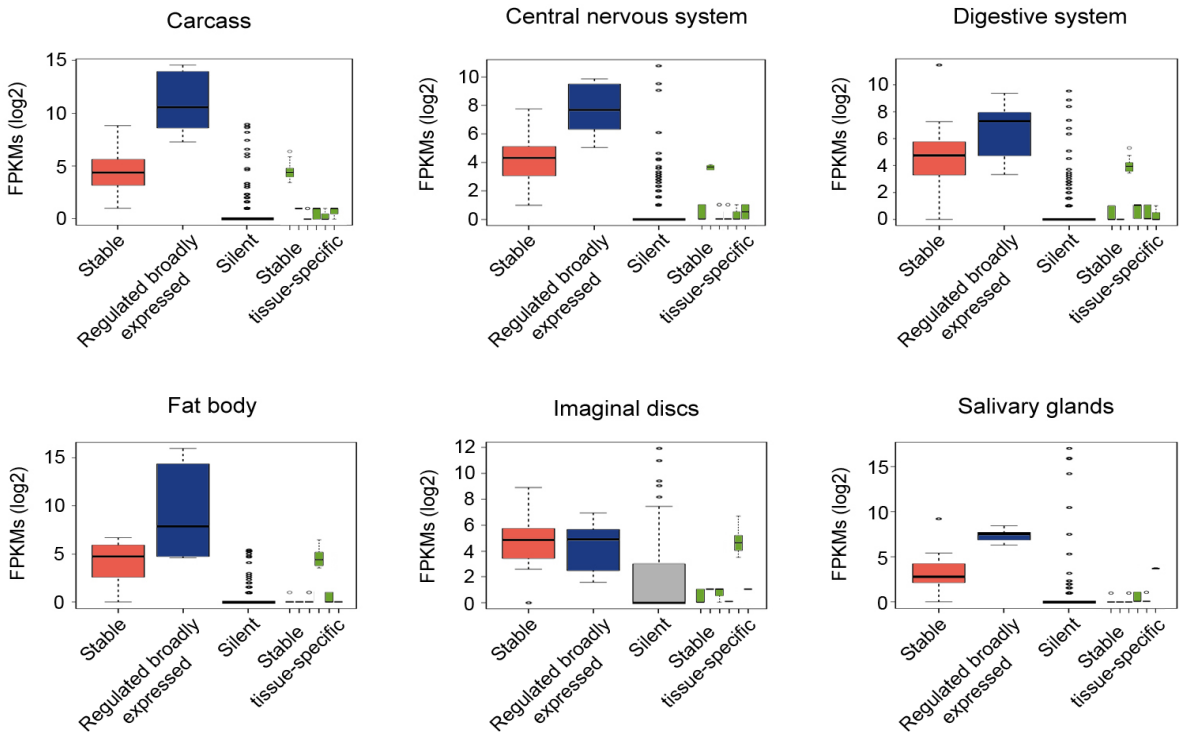
Supplementary Figure 3: Profiles of H3K4me3 along fly development for apterous (*ap*), a stable gene, and *Blastoderm-specific gene 25A* (*Bsg25A*), an embryo-specific gene. The expression (measured as FPKMs) along these points is given on the left. *In situ* hybridizations images obtained from BDGP[1] correspond to stages 13-16 (9-16h after egg laying) for *ap* and 1-6 (0-3h after egg laying) for *Bsg25A*, and show that these two genes have comparable restricted expression patterns.



Supplementary Figure 4: Partition of the entire set in stable and regulated genes. a, Distribution of the coefficient of variation on fly genes. The distribution of the coefficient of variation of gene expression along fly development reveals one major class of stable genes (P1), and two minor classes of genes that vary expression (P2 and P3). **b,** Number of genes belonging to each class. **c,** Distribution of gene expression levels at the developmental time point of maximum gene expression in each class. Gene expression is measured as FPKM by the modENCODE consortium **d,** Distribution of histone modifications at the time point of maximum gene expression in each gene class.



Supplementary Figure 5: Levels of H3K27me3 and H3K9me3 marking in stable and silent genes. **a,** Left panel: H3K27me3 in stable genes. As expected, most genes do not show H3K27me3. Right panel: H3K27me3 in silent genes. Many genes show either none or very low levels of H3K27me3. **b,** Left panel: H3K9me3 in stable genes. Most of genes do not show H3K9me3. Right panel: H3K9me3 in silent genes. Many genes show either none or very low levels of H3K9me3.



Supplementary Figure 6: Expression of stable genes, regulated genes broadly expressed at L3, silent genes, and stably expressed tissue-specific genes in six different tissues at L3. Expression levels, measured as FPKM by the modENCODE consortium, of six different tissues. The expression of stable tissue-specific genes is given for each tissue separately in the following order: Carcass, Central nervous system, Digestive system, Fat body, Imaginal discs, and Salivary glands. Regulated broadly expressed genes show higher expression than stable tissue-specific genes even in the tissue in which the later are expressed, except in Imaginal discs.

a

	WID cells		EID cells	
Total reads	133,777,031		111,712,574	
Genomic reads	108,639,465	(81.2%)	93,563,103	(83.8%)
Genomic reads. Unique	94,777,137	(70.8%)	86,740,797	(77.6%)
Intronic reads	1,014,566	(0.8%)	1,000,783	(0.9%)
Exonic reads	90,835,551	(67.9%)	83,346,225	(74.6%)
Intergenic read	2,927,020	(2.2%)	2,392,789	(2.14%)
Junction reads	15,057,579	(13.5%)	13,155,357	(11.8%)
Exons (RPKM>0)	50,610	(76.3%)	54,923	(82.8%)
Transcripts (RPKM >0)	17,657	(83.1%)	17,803	(83.8%)
Genes (RPKM >0)	11,507	(81.9%)	12,009	(85.4%)

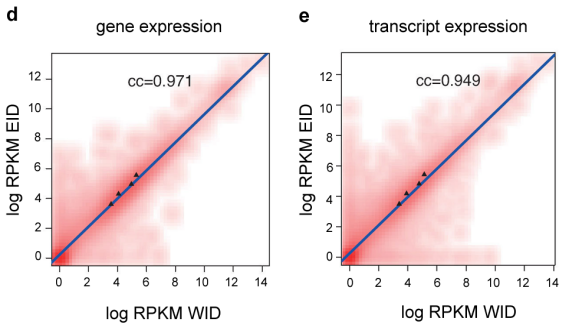
b

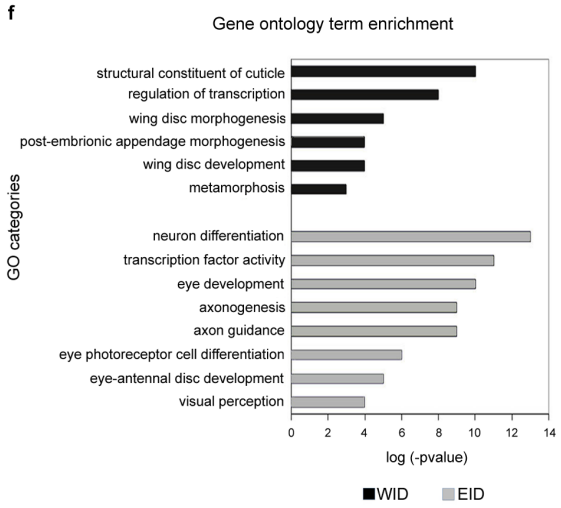
Gene RPKM cutoff	WID	EID	WID & EID
>0	11,507 (81.9%)	12,009 (85.4%)	11,167 (79.4%)
>1	7,502 (53.4%)	8,062 (57.3%)	7,373 (52.4%)
>2	7,098 (50.5%)	7,516 (53.5%)	6,975 (49.6%)
>4	6,666 (47.4%)	6,945 (49.4%)	6,513 (46.3%)
>8	5,916 (42.1%)	6,041 (43.0%)	5,667 (40.3%)
>16	4,425 (31.5%)	4,450 (31.7%)	4,072 (29.0%)
>32	2,545 (18.1%)	2,465 (17.5%)	2,203 (15.7%)

Transcript RPKM cutoff	WID	EID	WID & EID
>0	17,657 (83.1%)	17,803 (83.8%)	16,683 (78.6%)
>1	10,227 (48.1%)	11,124 (52.3%)	9,729 (45.8%)
>2	9,211 (43.3%)	9,930 (46.8%)	8,754 (41.2%)
>4	8,179 (38.5%)	8,625 (40.6%)	7,729 (36.4%)
>8	6,664 (31.4%)	6,987 (32.9%)	6,175 (29.0%)
>16	4,557(21.4%)	4,665 (21.9%)	4,067 (19.1%)
>32	2,447 (11.5%)	2,455 (11.6%)	2,093 (10.0%)

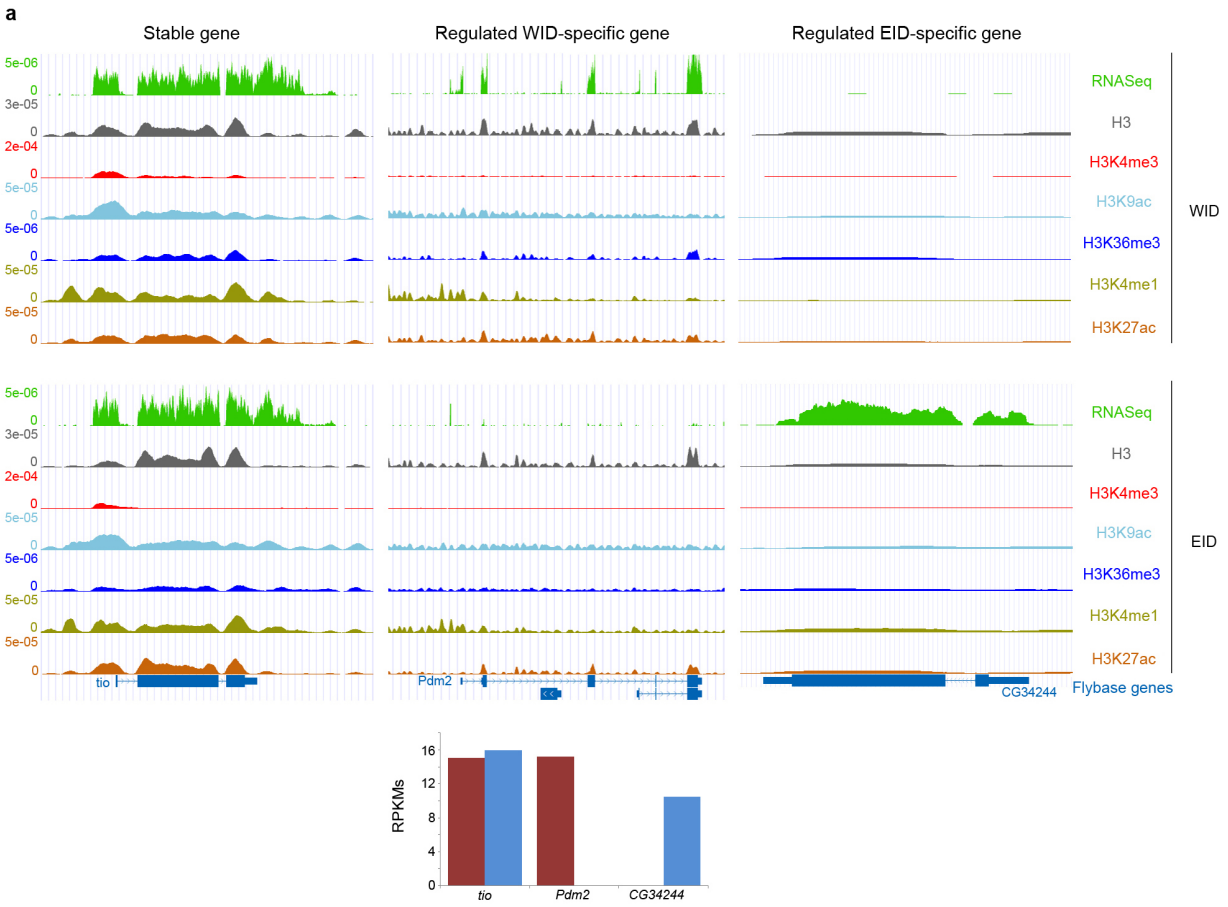
c

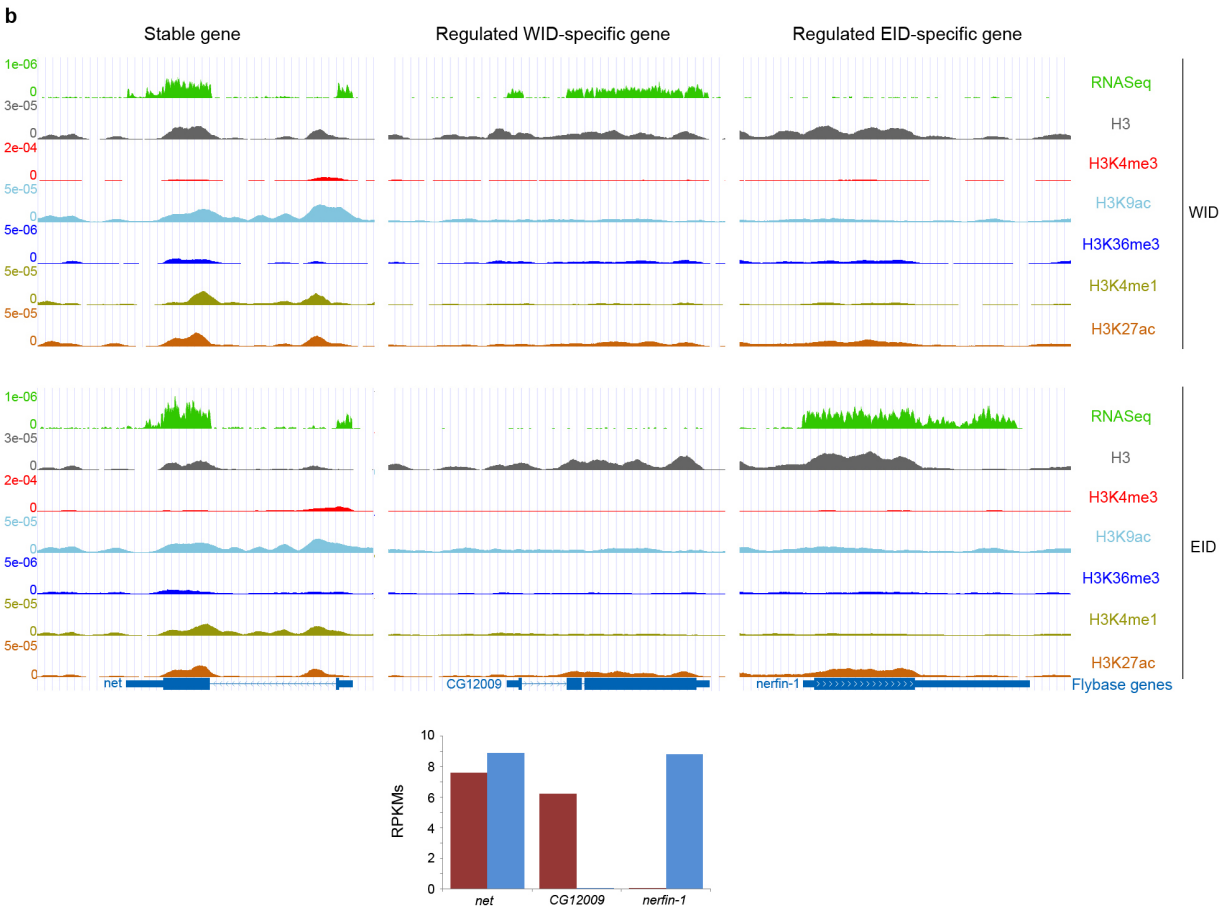
Tissue	Sample	Unambiguously mapped reads	Read length	Fragment length
WID	H3	9,504,688	36	317
	H3K4me3	4,389,032	36	325
	H3K9ac	34,855,962	50	206
	H3K36me3	9,611,956	36	325
	H3K4me1	26,742,570	36	310
	H3K27ac	20,352,560	36	304
EID	H3	16,183,101	36	326
	H3K4me3	27,809,050	36	325
	H3K9ac	44,523,481	50	180
	H3K36me3	10,580,356	36	345
	H3K4me1	11,218,055	40	305
	H3K27ac	18,467,364	36	302





Supplementary Figure 7: RNASeq and ChIPSeq analysis of Wing (WID) and Eye-antenna (EID) imaginal discs. a, RNASeq mapping and quantification statistics. Genomic reads are reads mapping to the genome. Genomic reads mapping uniquely are classified in three classes: intronic reads are reads mapping entirely within a gene, but not entirely within annotated exons. Exonic reads are reads mapping entirely within exons. Intergenic reads are reads not mapping entirely within genes. Junction reads are reads mapping to splice junctions but not to the genome. **b,** Number of genes and transcripts expressed at different expression cutoffs. **c,** Mapping statistics for the ChIPSeq experiments on histone modifications. The genome-wide Pearson correlation between WID and EID epigenomes is very high: 0.90 for H3, 0.84 for H3K4me3, 0.94 for H3K9ac, 0.96 for H3K36me3, 0.92 for H3K4me1 and 0.92 for H3K27ac when computed on the number of reads mapping onto 1,000 bp long windows. **d, e,** Join distribution in WID and EID of gene and transcript expression. Expression is measured in log RPKM. **f**, Gene Ontology term enrichment of 628 genes preferentially expressed in EID and 184 genes preferentially expressed in WID.


a
Stable gene
Regulated WID-specific gene
Regulated EID-specific gene
RNASeq
H3
H3K4me3
H3K9ac
H3K36me3
H3K4me1
H3K27ac
WID
EID
tio
Pdm2
CG34244
Flybase genes
RPKMs
tio
Pdm2
CG34244


b
Stable gene
Regulated WID-specific gene
Regulated EID-specific gene
RNASeq
H3
H3K4me3
H3K9ac
H3K36me3
H3K4me1
H3K27ac
WID
EID
net
CG12009
nerfin-1
Flybase genes
RPKMs
net
CG12009
nerfin-1

Supplementary Figure 8: Profiles of RNA expression, H3 and histone modifications in WID and EID-specific genes. a, Stable gene *tio*, regulated WID-specific gene *Pdm2* and regulated EID-specific gene *CG34244* are expressed at the same level (green tracks and bottom panel). Histone modifications typical of gene activation are observed in *tio* whereas the tissue-specific genes lack all of them, even in the tissue in which they are expressed. **b,,** Stable gene *net*, regulated WID-specific gene *CG12009* and regulated EID-specific gene *nerfin-1* are expressed at very similar levels (green tracks and bottom panel), but *net* exhibits histone modifications, whereas the tissue-specific genes lack all of them, even in the tissue in which they are expressed. In none of the cases, absence of histone marking cannot be attributed to the lack of nucleosomes because the H3 signal is comparable in both cases.

Supplementary Table 1: Gene and Exon sets.

Set	Condition	WID	EID
GENES			
Expressed	RPKM > 0	11,507	12,009
Differentially expressed	Genes lying at least 1 unit above the regression line	184	628
Regulated tissue-specific	RPKM >= 1.5 in one tissue and < 0.1 in the other tissue and cv >= 1.2	10	55
Stable expressed in WID and EID	RPKM > 2.3 in the two tissues and within 20% difference and cv < 1.2	284	
Silent in WID and EID	RPKM = 0 in the two tissues and cv < 1.2	30	
EXONS			
Highly included	Inclusion > 0.9	744	821
Lowly included	Inclusion < 0.1	104	97

Supplementary Table 2a: Correlation between transcriptional stability (measured as coefficient of variation, cv), histone modification levels and gene expression levels. For each gene, we compute the cv along the twelve fly developmental time points, and the average expression and histone modification levels. The correlation is computed over all genes.

	Stable genes (cv < 1.2)				All genes			
	cv *vs.* histone levels		Histone levels *vs.* gene expression		cv *vs.* histone levels		Histone levels *vs.* gene expression	
mark	CC	p-value	CC	p-value	CC	p-value	CC	p-value
H3K4me3	-0.4	< 2.2e-16	0.4	< 2.2e-16	-0.67	< 2.2e-16	0.28	< 2.2e-16
H3K9ac	-0.34	< 2.2e-16	0.43	< 2.2e-16	-0.62	< 2.2e-16	0.26	< 2.2e-16
H3K4me1	-0.21	< 2.2e-16	0.24	< 2.2e-16	-0.52	< 2.2e-16	0.21	< 2.2e-16
H3K27ac	-0.33	< 2.2e-16	0.42	< 2.2e-16	-0.63	< 2.2e-16	0.31	< 2.2e-16

Supplementary Table 2b: Partial correlations between transcriptional stability (measured as coefficient of variation, cv) and histone modification levels, eliminating the effect of gene expression, For each gene, we compute the cv along the twelve fly developmental time points, and the average expression and histone modficiation levels. To compute the partial correlation between cv and average histone modifications, we used the Package GGM, giving the gene expression as controlling variable.

	cv *vs*. histone levels (controlling by gene expression)			
	Stable genes (cv < 1.2)		All genes	
mark	CC	p-value	CC	p-value
H3K4me3	-0.34	< 2.2e-16	-0.68	< 2.2e-16
H3K9ac	-0.27	< 2.2e-16	-0.63	< 2.2e-16
H3K4me1	-0.17	< 2.2e-16	-0.52	< 2.2e-16
H3K27ac	-0.26	< 2.2e-16	-0.64	< 2.2e-16

Supplementary references

1. Hammonds, A.S. et al. Spatial expression of transcription factors in Drosophila embryonic organ development. *Genome Biol* **14**, R140 (2013).

Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs.

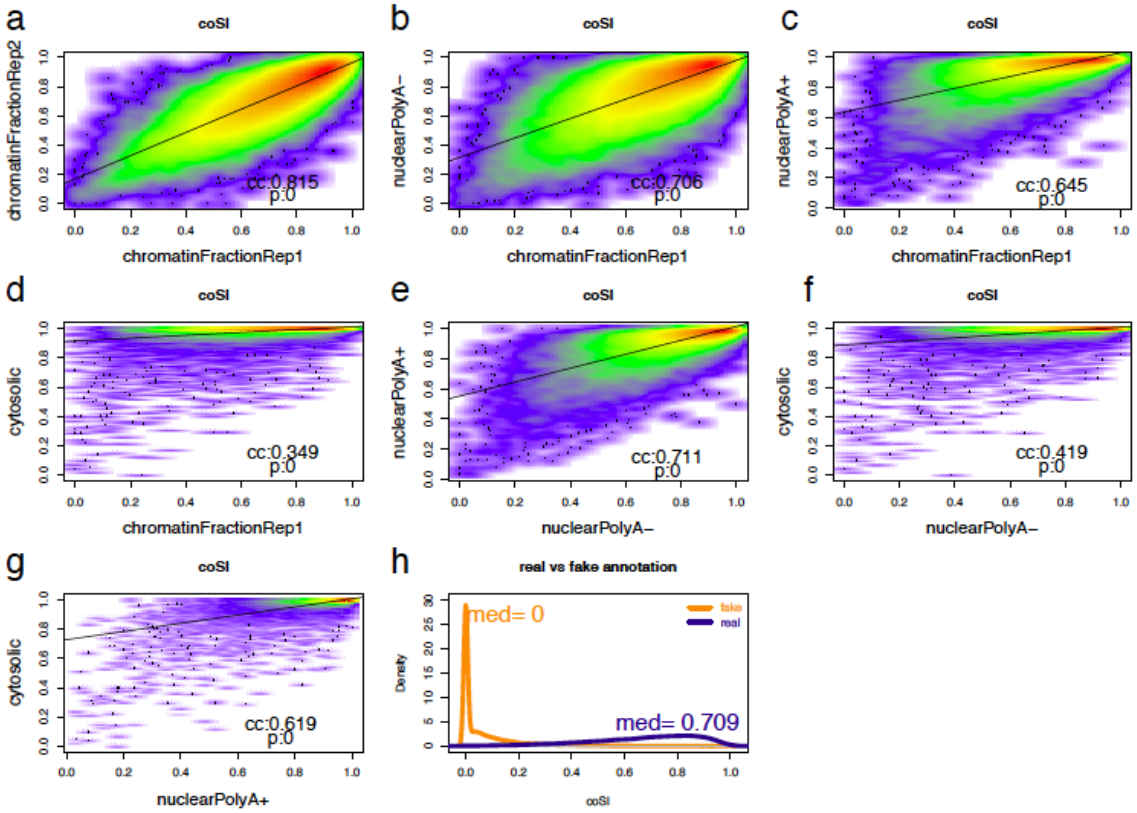


Figure S1: Correlation analysis of coSI values between different subcellular RNA fractions using exons that had high enough read numbers in all these sub-cellular fractions. Correlation between two replicas of the chromatin associated total RNA fraction **(a)**, the total chromatin associated RNA fraction and the polyA- nuclear fraction **(b)**, the total chromatin associated RNA fraction and the polyA+ nuclear fraction **(c)**, the total chromatin associated RNA fraction and the polyA+ cytosolic fraction **(d)**, the polyA- nuclear fraction and the polyA+ nuclear fraction **(e)**, the polyA- nuclear fraction and the polyA+ cytosolic fraction **(f)**, the polyA+ nuclear fraction and the polyA+ cytosolic fraction **(g)**. Comparison of coSI values using the real annotation and a fake annotation (shifted by 30bps) against transcription direction **(h)**.

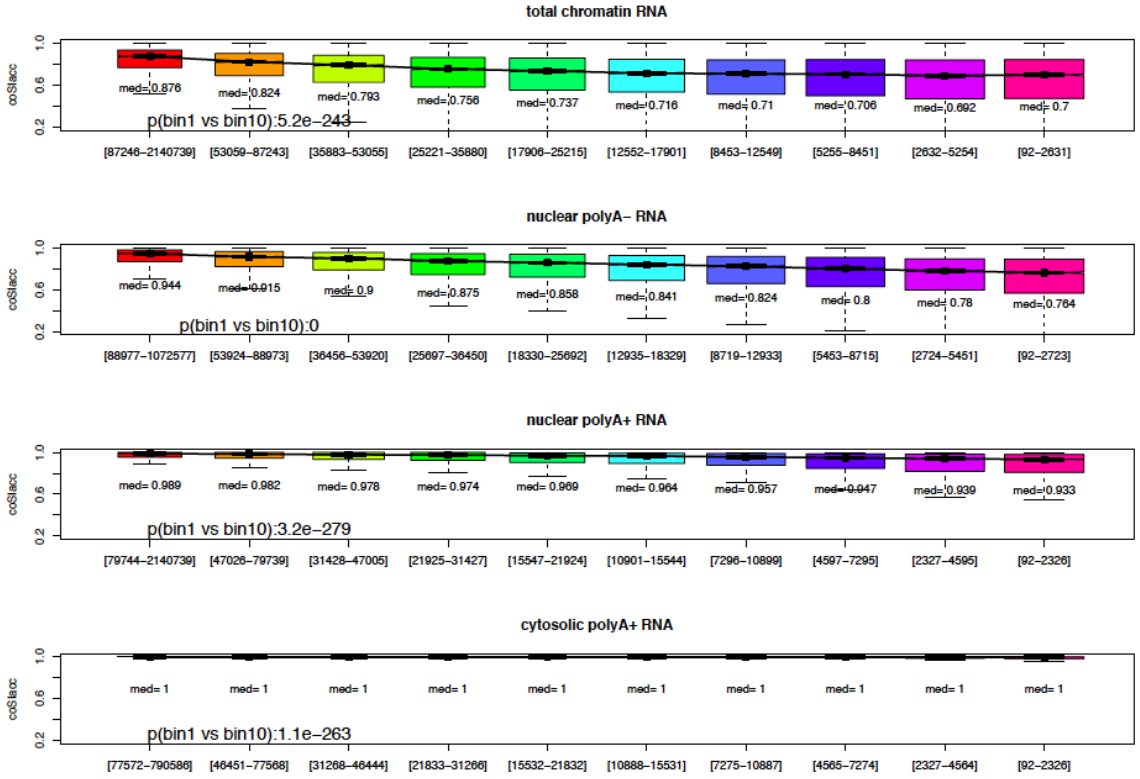


Figure S2: Boxplots of acceptor based coSI values (formula: a/(a+c) in the terminology of figure 1b) in bins according to the distance of an exon to the annotated polyA-site for the total chromatin associated RNA fraction (a) the polyA- nuclear fraction (b) the polyA+ nuclear fraction (c) and the polyA+ cytosolic fraction (d). P-values were calculated comparing the first and the last bin, using a two-sided Wilcoxon rank sum test.

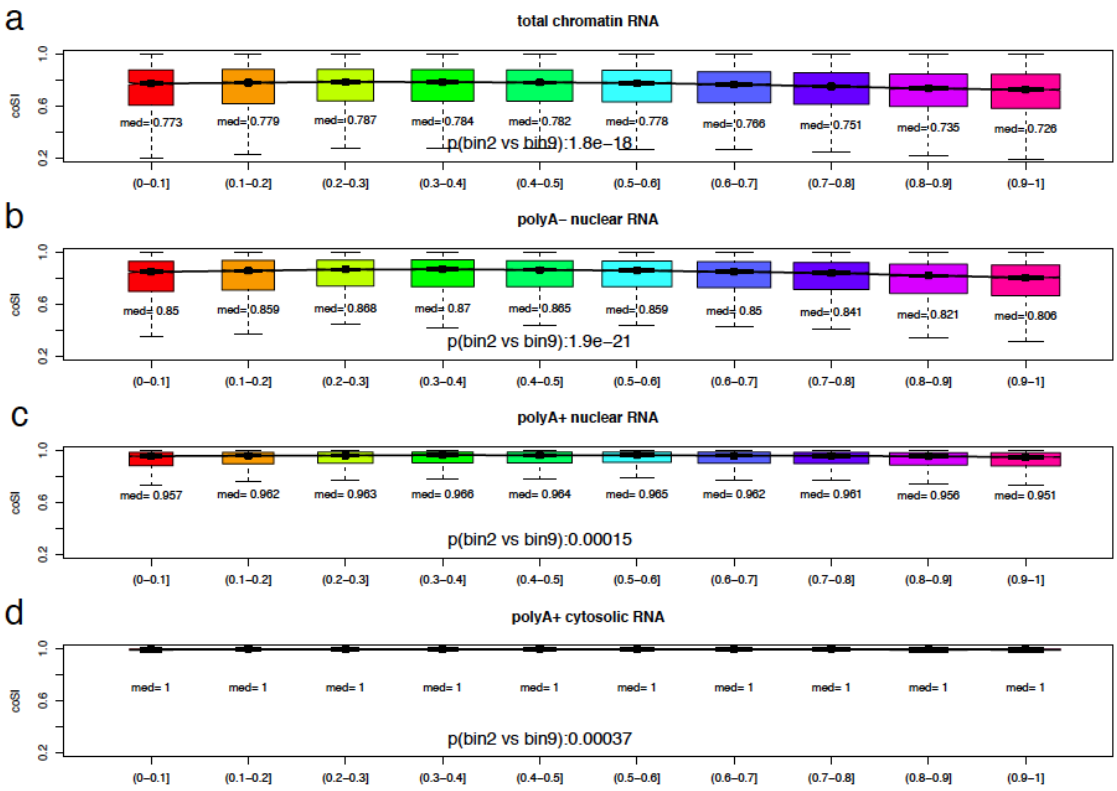


Figure S3: Boxplots of coSI value in bins according to the relative position of an exon in the gene (intervals on x-axis give minimum and maximum percentage with respect to the total gene length) for the total chromatin associated RNA fraction (a) the polyA- nuclear fraction (b) the polyA+ nuclear fraction (c) and the polyA+ cytosolic fraction (d). P-values were calculated comparing the second and the last but one bin, using a two-sided Wilcoxon rank sum test.

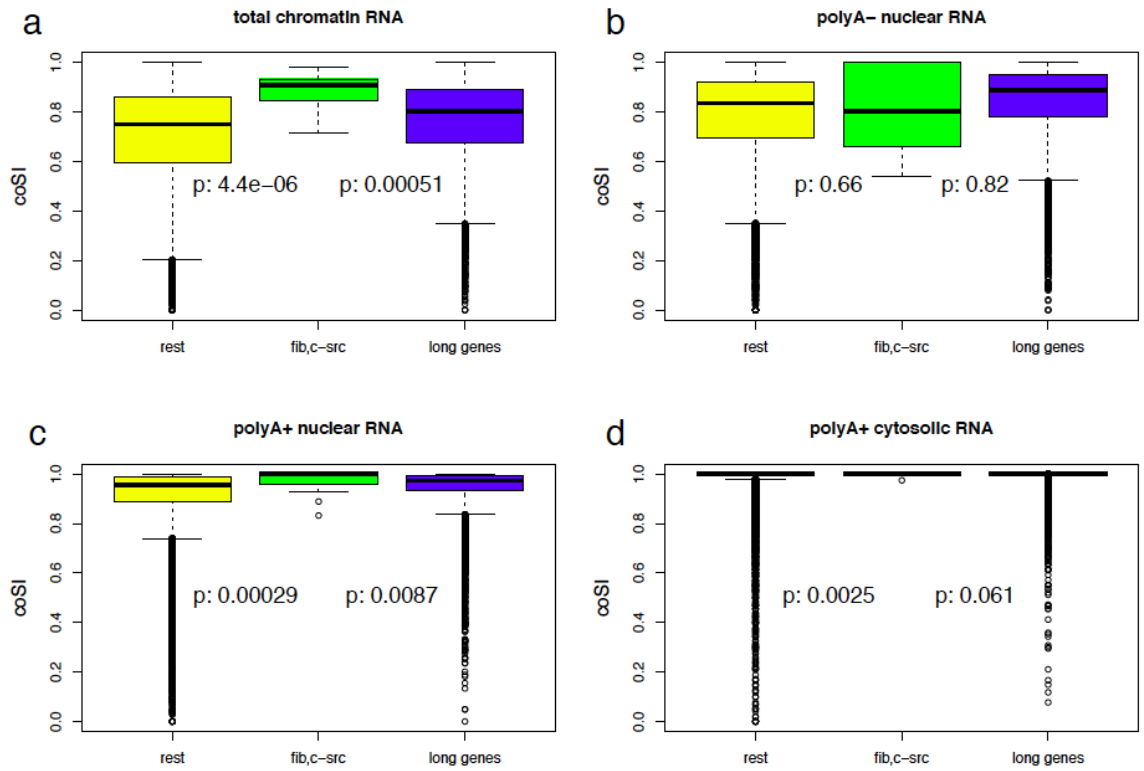


Figure S4: Boxplots of coSI values of exons for genes other than fibronectin and c-src (yellow), of exons for the fibronectin and c-src gene (green) and for the 4000 longest genes in the annotation (blue), which exclude fibronectin and c-src; the total chromatin associated RNA fraction (a) the polyA- nuclear fraction (note, that in this fraction, we could calculate the coSI value for only 8 exons due to lower read numbers, b) the polyA+ nuclear fraction (c) and the polyA+ cytosolic fraction (d). P-values were calculated using a one-sided Wilcoxon rank sum test.

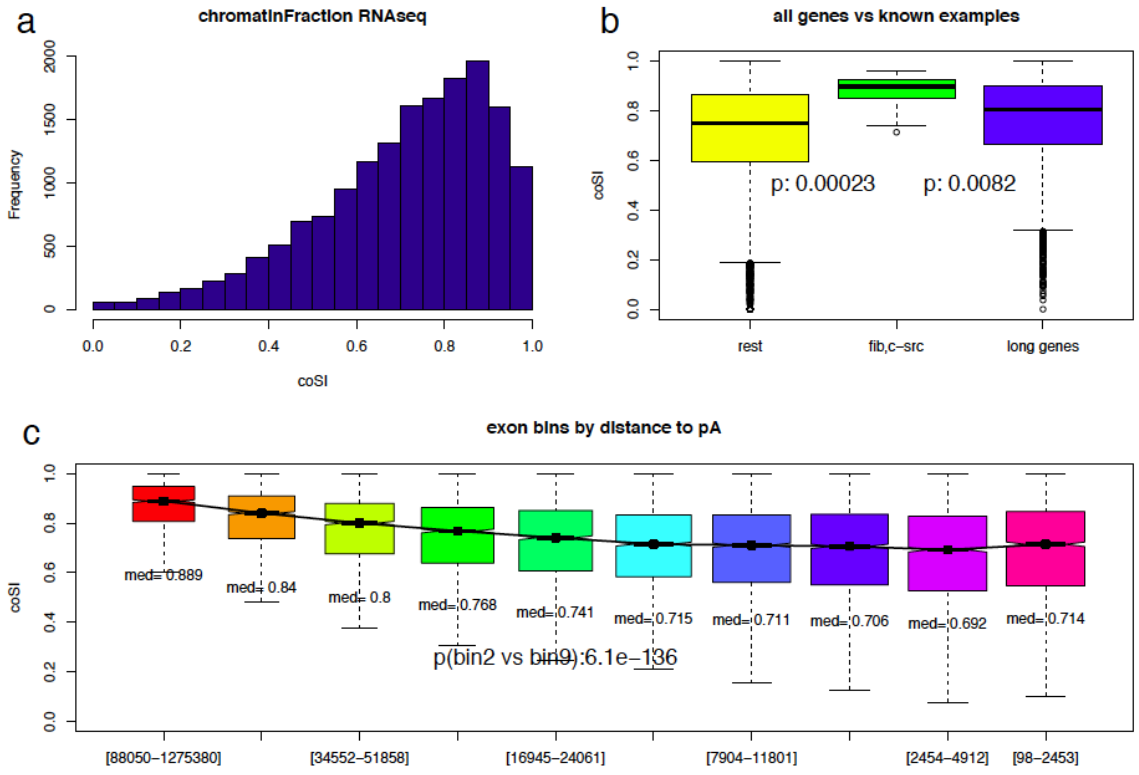


Figure S5: Analysis of total chromatin RNA-seq when excluding all exons that have any potential novel splice site (as indicated by split mappings of previously unmapped reads). Histogram of coSI values (a). Boxplots of coSI values of exons for genes other than fibronectin and c-src (yellow), of exons for the fibronectin and c-src gene (green) and for the 4000 longest genes in the annotation (blue), which exclude fibronectin and c-src (b). Boxplots of coSI values in bins according to the distance of an exon to the annotated polyA-site (Intervals on x-axis give minimum and maximum distance in each bin) (c).

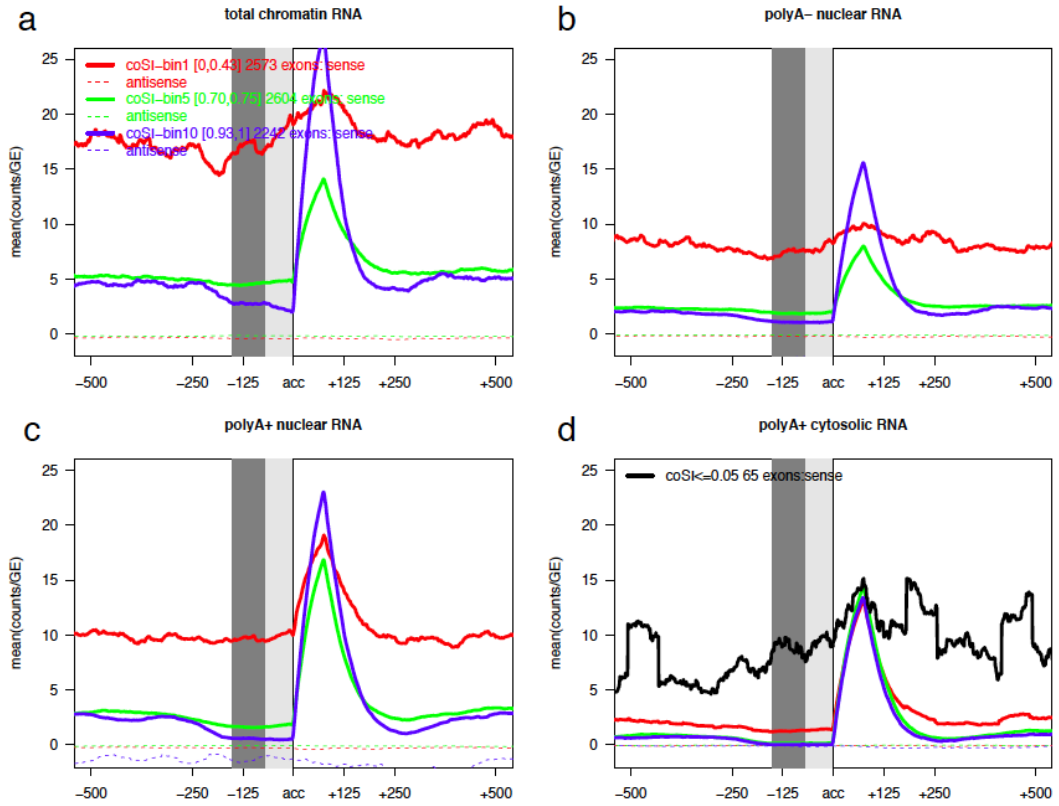


Figure S6: Similar plot to figure 3, but normalised for cytosolic polyA+ gene expression. In addition we show the cytosolic profile of exons with extremely coSI values in the total chromatin fraction. This profile is much flatter, supposedly because this exon set contains many exons around which introns are retained (d).

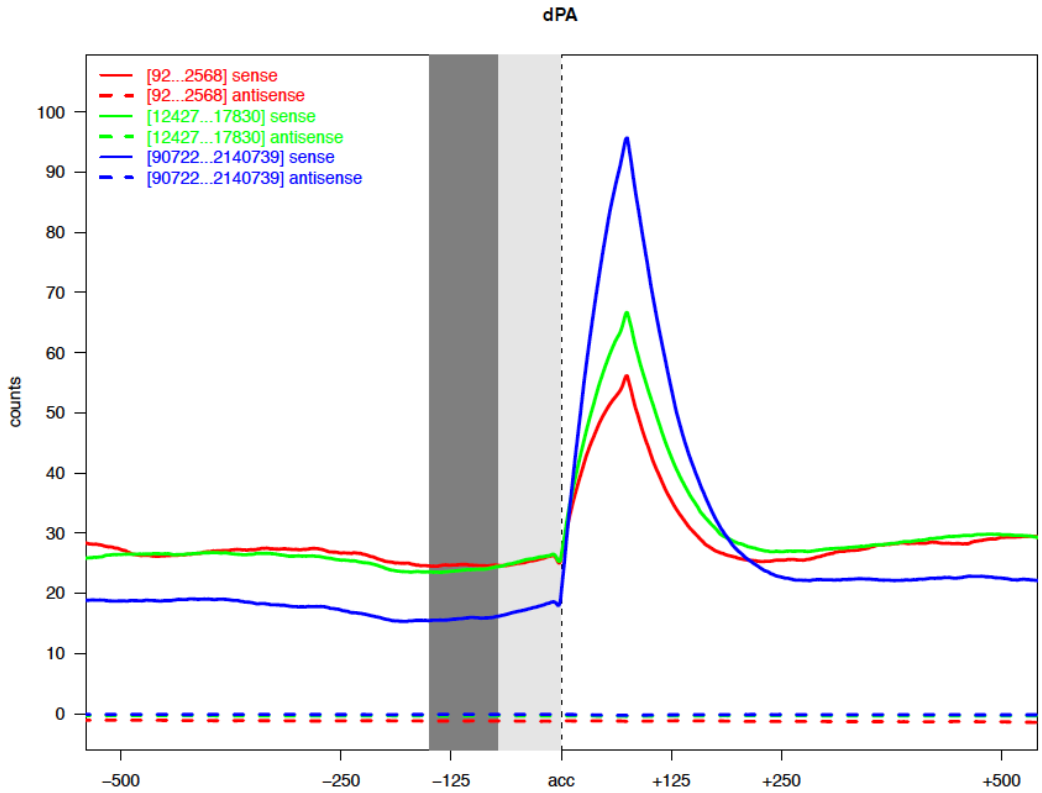


Figure S7: Similar plot to figure 3 with exon binning by distance to polyA-site, showing that exons far from the polyA-site have higher exonic coverage but lower intronic coverage than exons close to the polyA-site. Reads mapping deep within exons or introns, that were not used for the coSI calculation therefore show a similar trend of lower completed splicing as one approaches the polyA-site. Legend gives minimal and maximal distance to the polyA-site in each exon-bin.

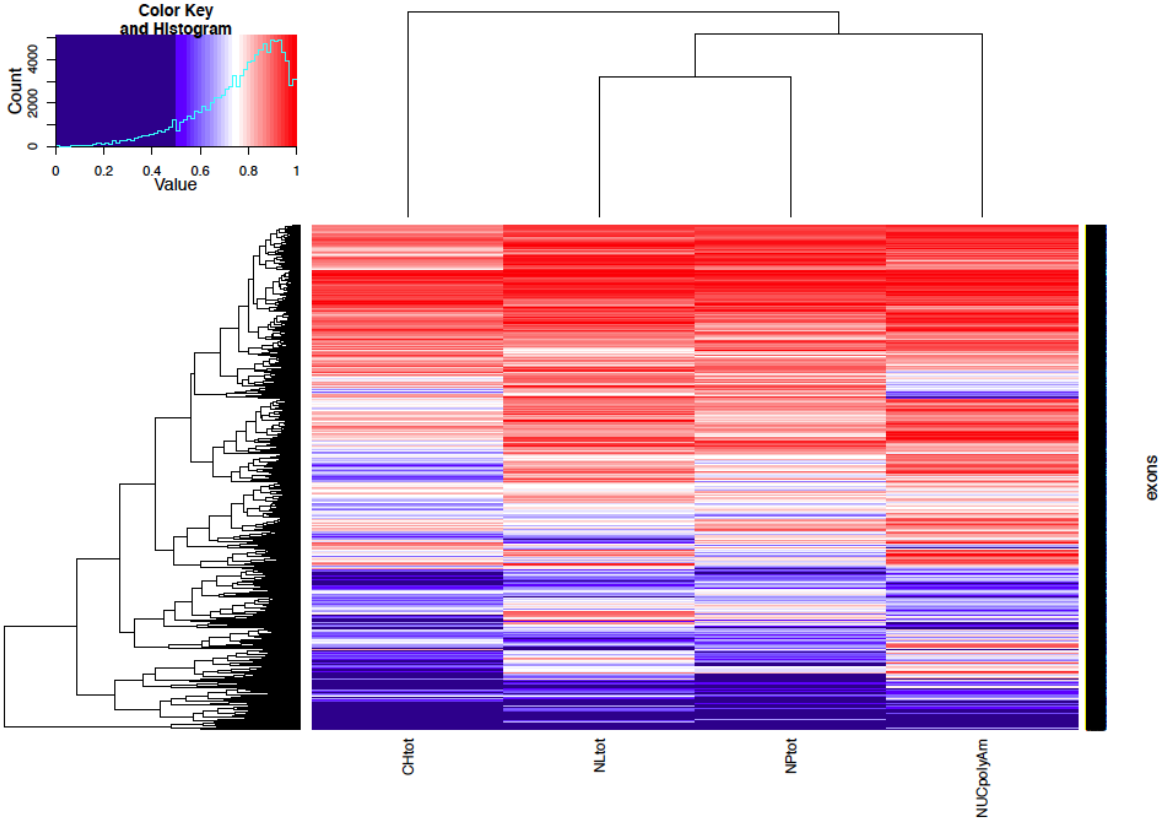


Figure S8: Clustering of sub-cellular RNA fractions and exons according to exonic coSI values using four RNA fractions (from left to right): total chromatin associated RNA, total nucleoli RNA, total nucleoplasm RNA, polyA- nuclear RNA. The total nucleoli and nucleoplasm fractions clusters with polyA- nuclear RNA before clustering with total chromatin RNA, showing that the experiment type (total RNA vs. polyA- RNA) does not determine the clustering. Note that the color scale is linear only from coSI¿=0.5 on.

compartment	*cov(intron)%*	*cov(exon)%*
total chromatin	42.3	59.1
nuclear polyA-	38.0	56.1
nuclear polyA+	31.3	59.1
cytosolic polyA+	6.0	49.4
cytosolic polyA-	1.8	31.9
total nucleolus	31.5	51.4
total nucleoplasm	42.6	57.9

Table S2: Coverage analysis. Sub-cellular compartment (column 1). Percentage of intronic nucleotides covered by long RNA-seq reads (column 2). Percentage of exonic nucleotides covered by long RNA-seq reads (column 3).

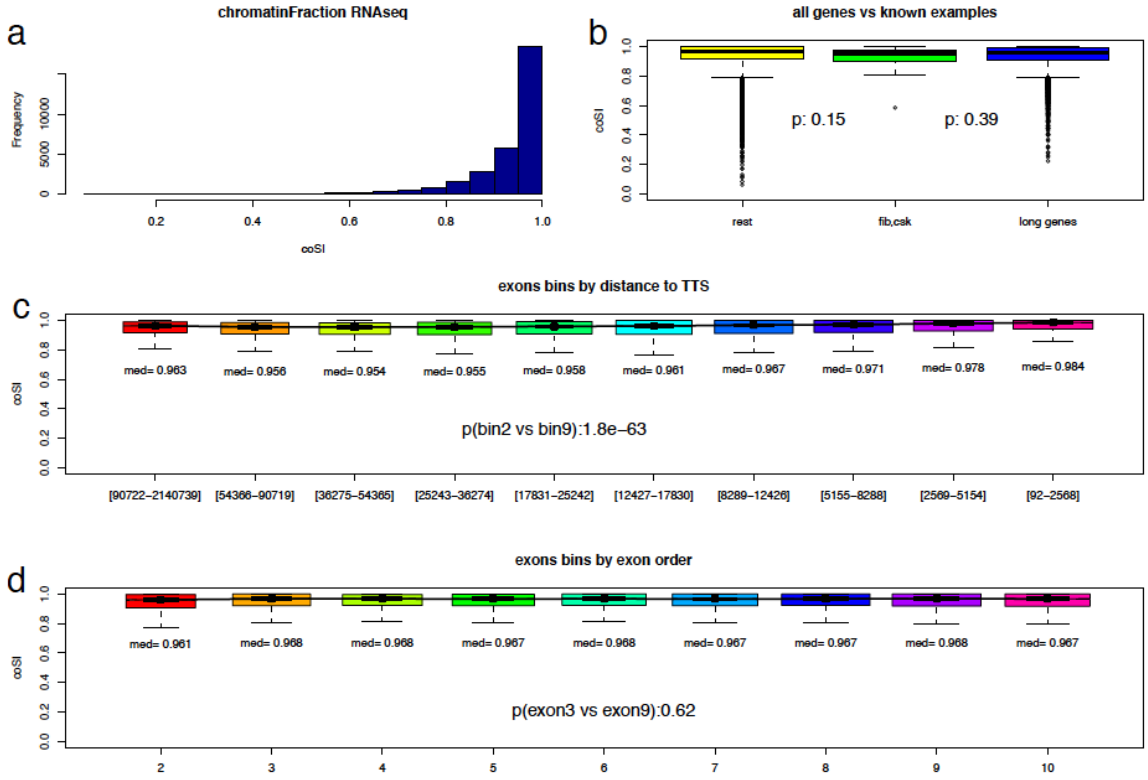


Figure S9: coSI-CAGE calculation: unspliced reads were replaced by CAGE reads in the total chromatin fraction. Histogram of coSI-CAGE values in the total chromatin RNA fraction (a). Boxplots of coSI-CAGE values of exons for genes other than fibronectin and c-src (yellow), of exons for the fibronectin and c-src gene (green) and for the 4000 longest genes in the annotation (blue), which exclude fibronectin and c-src (b). Boxplots of coSI-CAGE values in bins according to the distance of an exon to the annotated polyA-site (c). Boxplots of coSI-CAGE values in bins according to exon order within the transcripts (d). The trends that we had observed in coSI values are either not significant (b) or inversed (c,d) when considering coSI-CAGE values. This shows that (1) unspliced reads do not originate from novel, unidentified TSS and (2) that observations made for the coSI are not an artifact of any, random sequencing experiment.

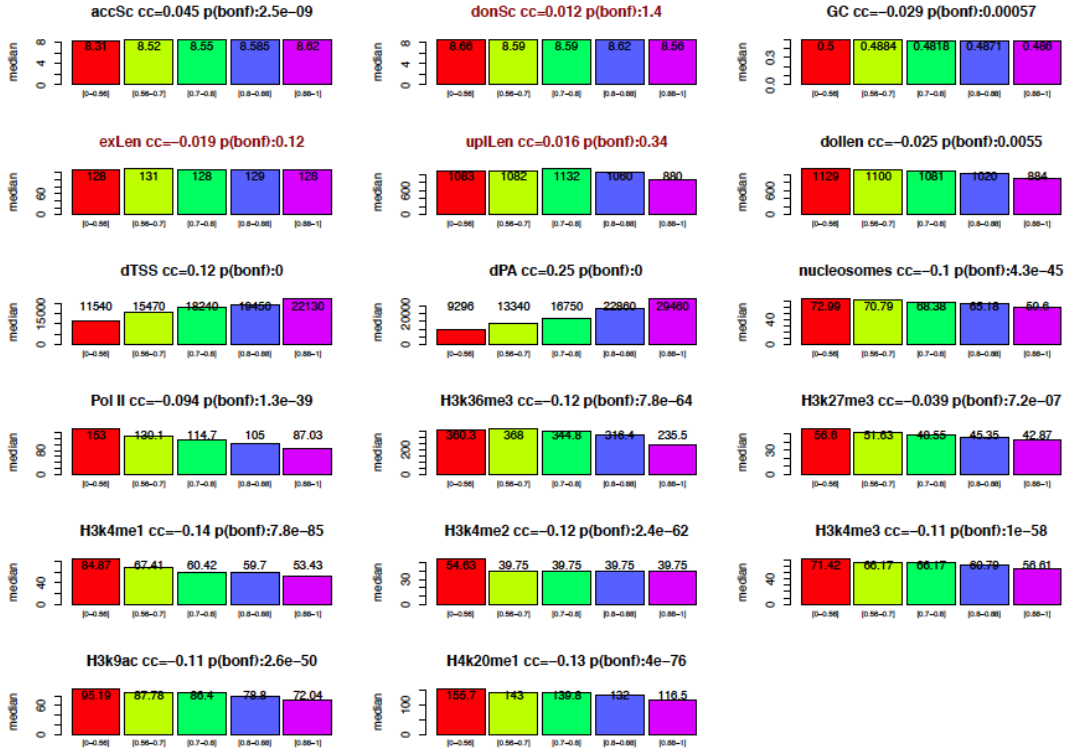


Figure S10: Raw correlation between gene, exon and chromatin structure variables on the one hand and coSI values on the other hand. Median values for each variable are depicted in five bins of coSI values. Investigated variables: acceptor strength (accSc), donor strength (donSc), exonic GC content (GC), exon length (exLen), upstream intron length (upILen), downstream intron length (doIlen), distance to the TSS (dTSS), distance to the polyA-site (dPA), MNase level on first 147bps after acceptor (nuc), same for Pol II (Pol II), same for 5 histone modifications HXkYmeZ (HXkYmeZ).

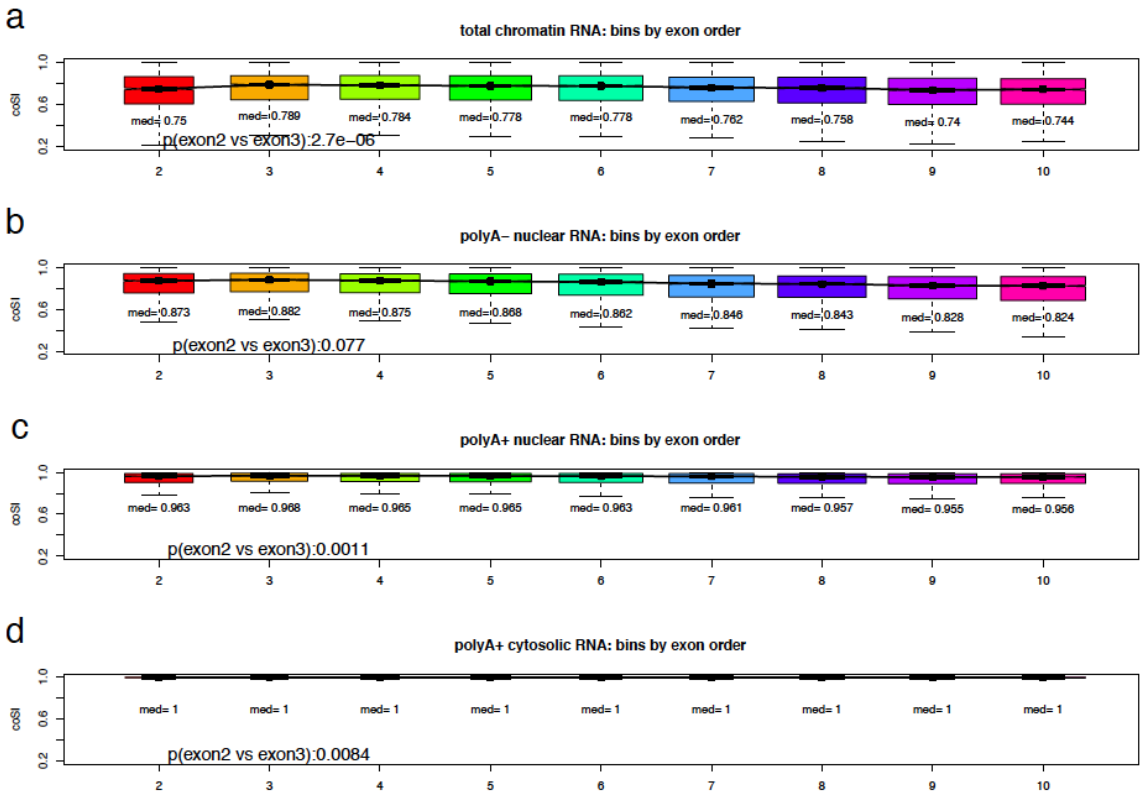


Figure S11: Boxplots of coSI value in bins according to exon order within a transcript for the total chromatin associated RNA fraction (a) the polyA- nuclear fraction (b) the polyA+ nuclear fraction (c) and the polyA+ cytosolic fraction (d). P-values were calculated comparing the 2nd and 3rd exon, using a two-sided Wilcoxon rank sum test.

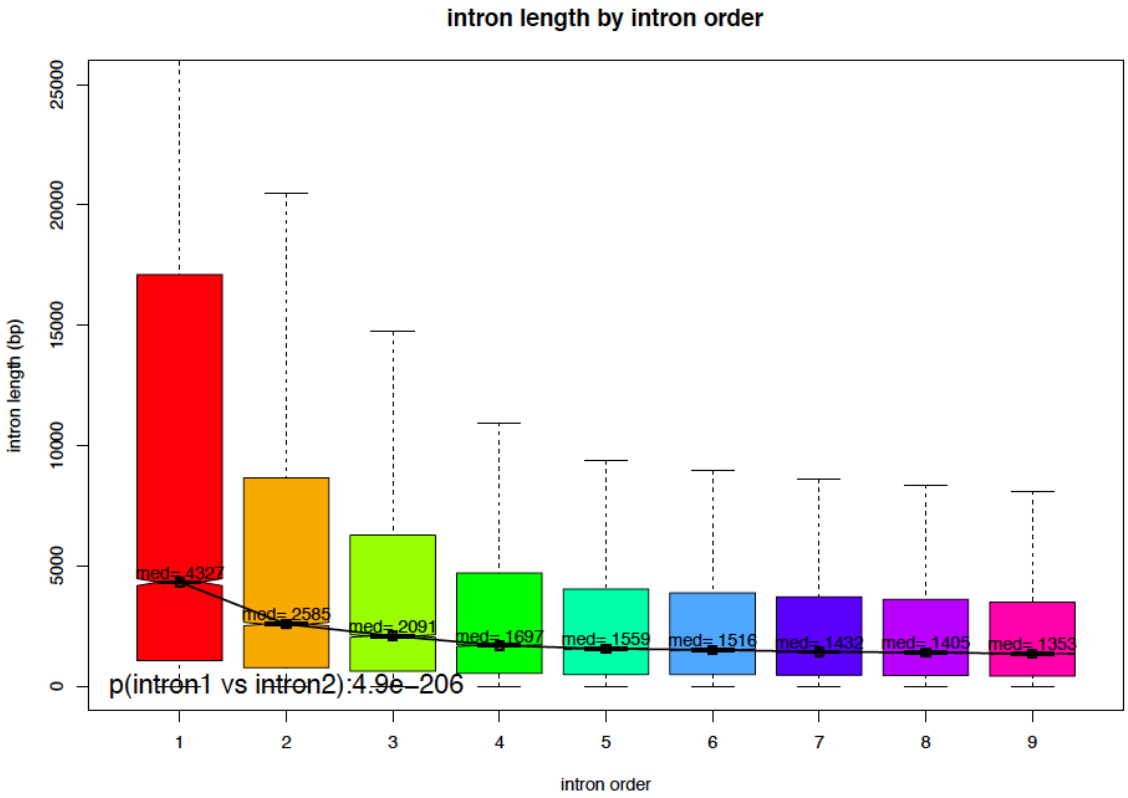


Figure S12: Boxplots of intron length as a function of intron order. In this analysis an intron was counted n times if it appeared in n different transcripts. Only transcripts with at least 10 exons were used.

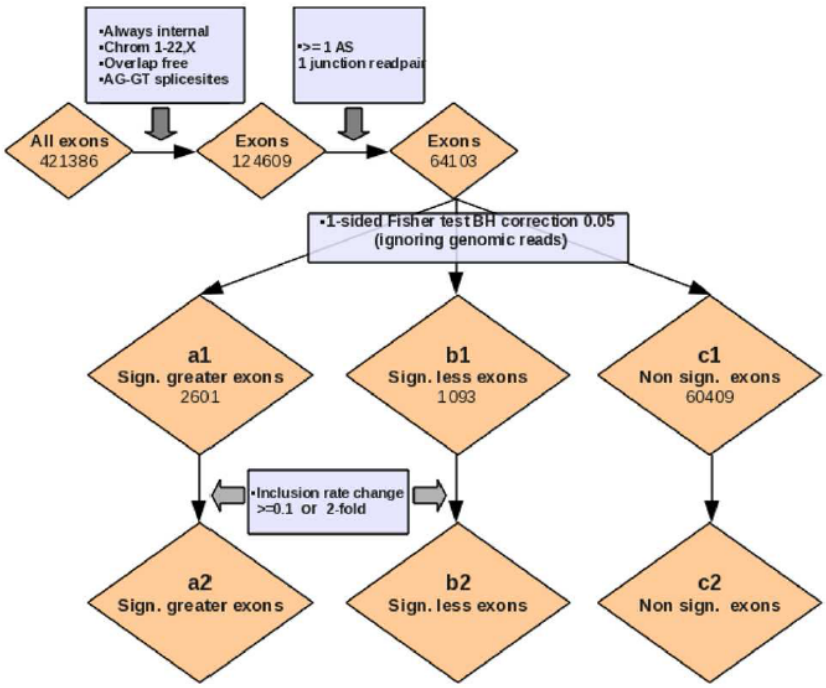


Figure S13: Diagram describing the pipeline to call alternative exons in four pairwise cell type comparisons (K562 on the one hand and Gm12878,Hepg2, HelaS3 and Huvec on the other hand). Orange diamonds: exon sets: Purple rectangles: operations.

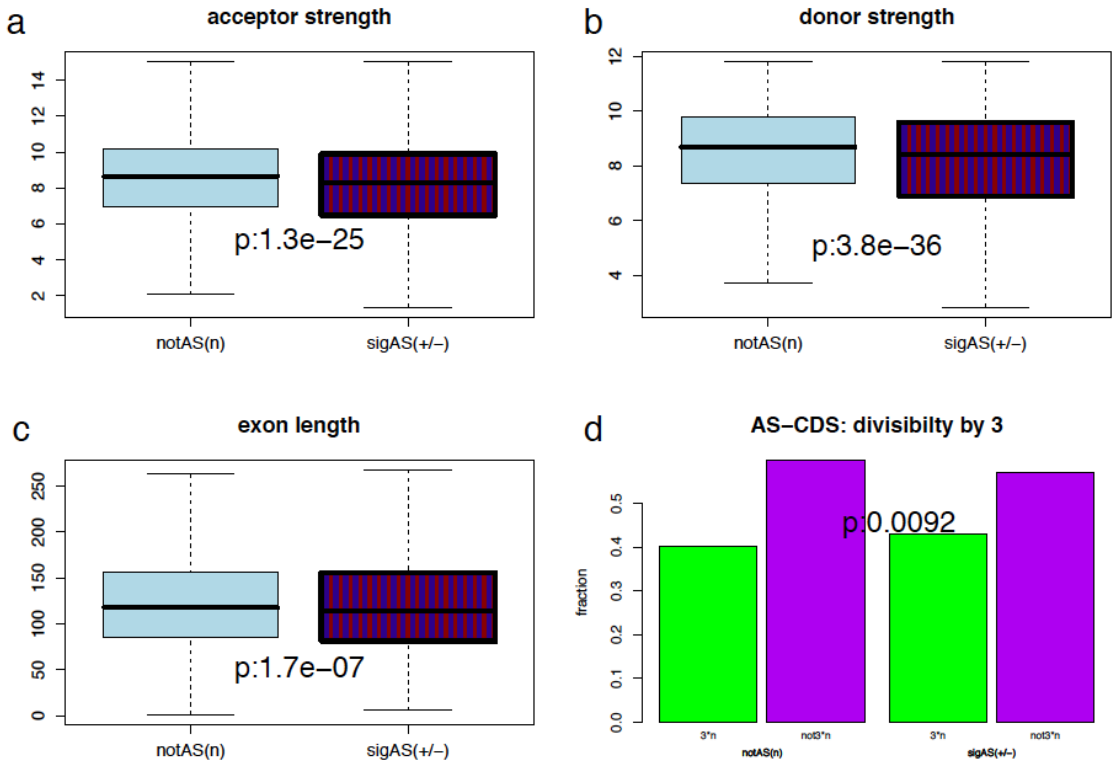
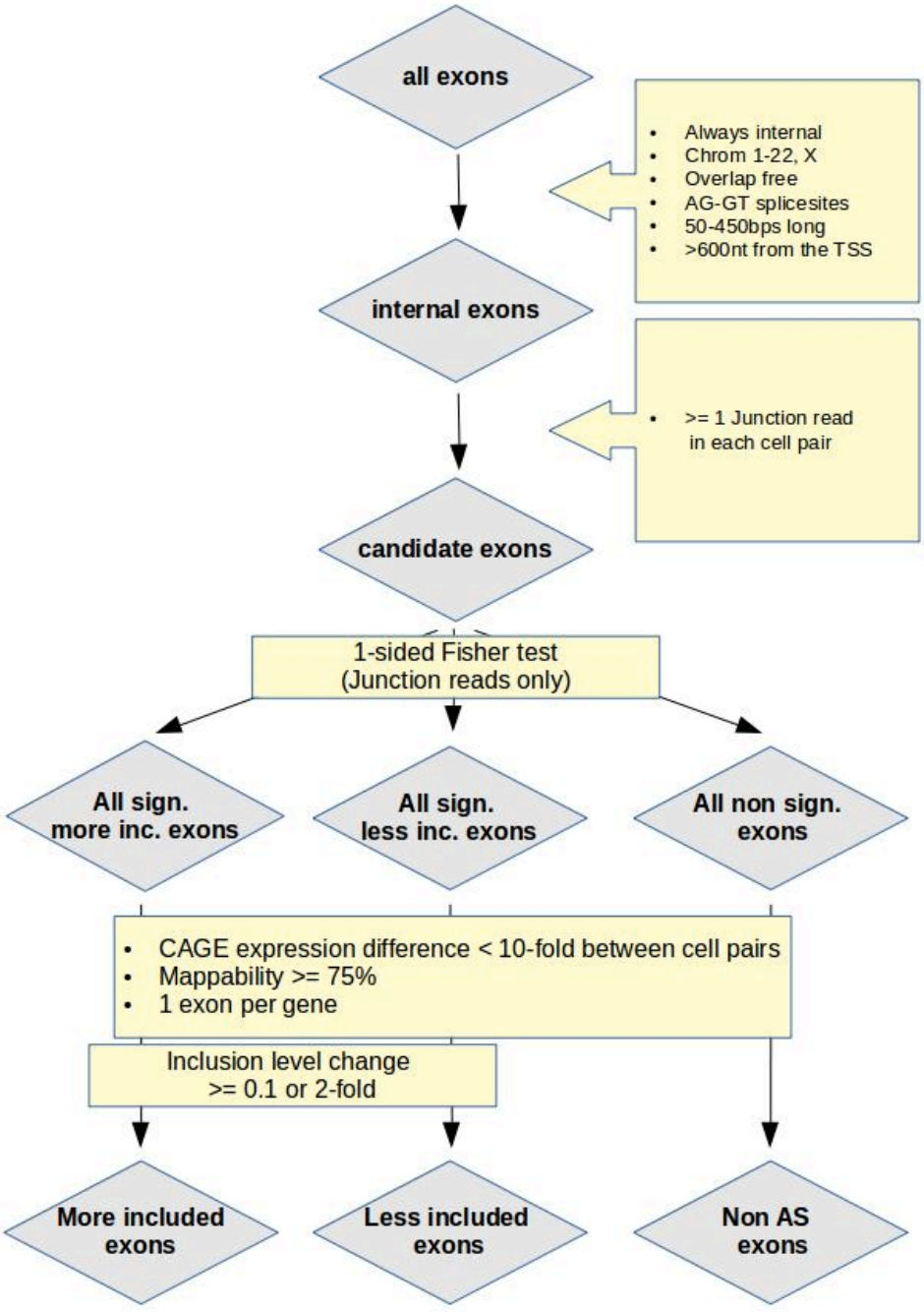


Figure S14: Characteristics of exons that were determined as having cell type specific inclusion levels as compared to those that did not show cell type specific inclusion levels in any of the four pairwise cell type comparisons: acceptor strength (a), donor strength (b), exon length (c), CDS-divsibility-by-3 (d).

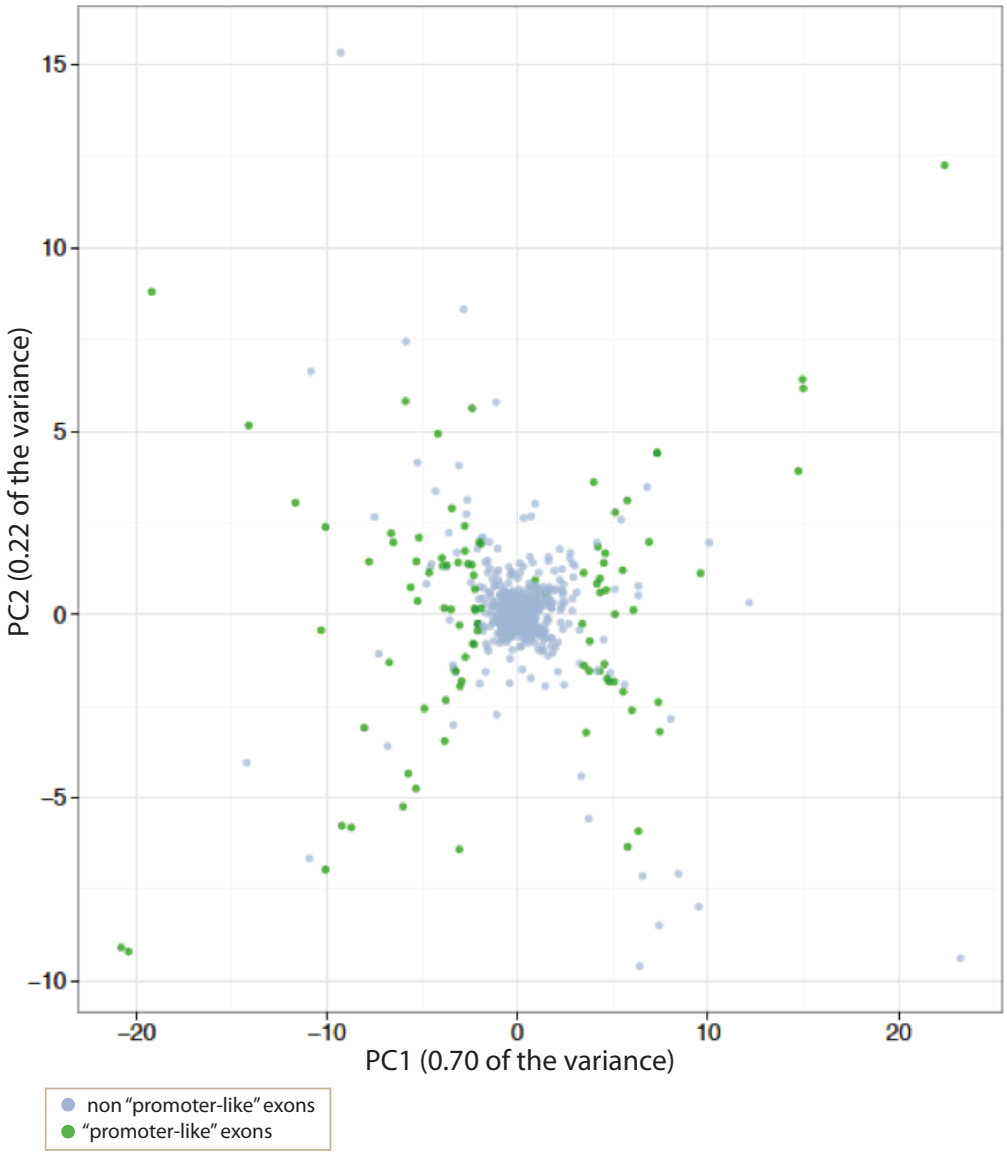
Promoter-like epigenetic signatures in exons with cell type-specific splicing

Figure S1. Flowchart of the pipeline employed to define differentially included exons between two cell lines



Internal exons were selected based on Gencode 15 annotation. Junction reads from RNASeq datasets of the cell types investigated within the ENCODE project were used to define the lists of alternative (more and less included) and not alternative (notAS) exons for each cell pair. These exons were also filtered by expression level changes (CAGE data), mappability, inclusion level change and only one per gene was kept.

Figure S2. Principal component analysis (PCA) of the differentially included exons based on the differential H3K9ac, H3K27ac and H3K4me3 ChIPSeq signal.

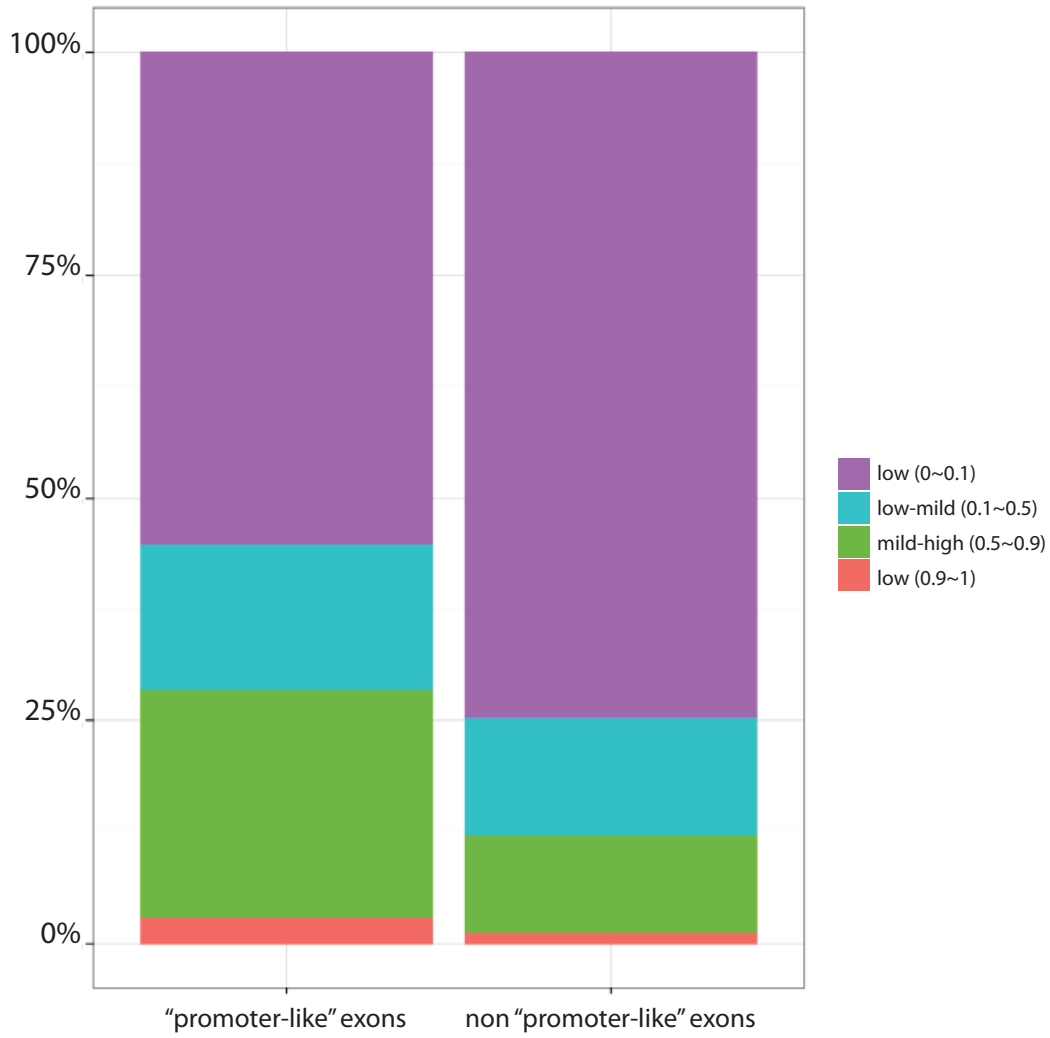


"Promoter-like" exons are represented in green and the remaining differentially included exons in gray. The left and right subgroups of "promoter-like" exons, separated by PC1, represent the less and more included exons, respectively.

Figure S3. Differential inclusion and chromatin levels between K562 and NHEK

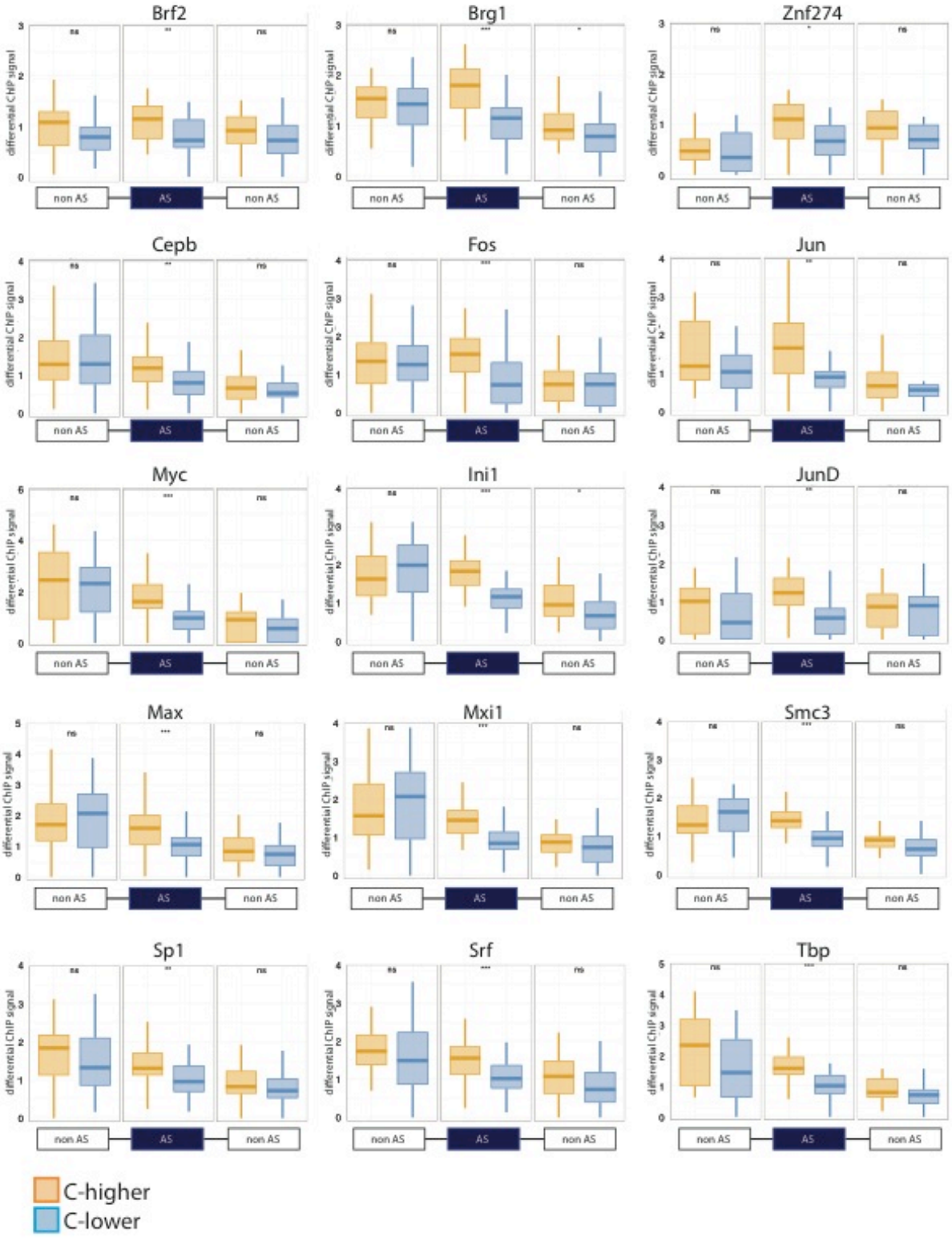
In the 22 “promoter-like” exons that were regulated between K562 and NHEK we calculated the differential inclusion level (in the plot multiplied by 100), the differential signal of the H3K9ac, H3K27ac and H3K4me3 levels and the Total of the three. In the overwhelmingly majority of the cases, differential exon inclusion and differential total histone levels are in consistent directions.

Figure S4. Inclusion levels distribution in tissue samples from GTEx.



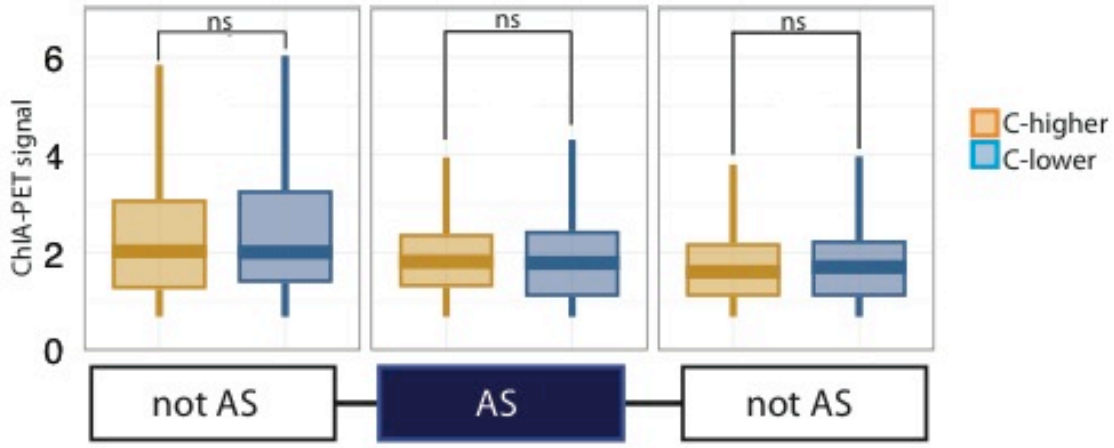
Inclusion levels for all the identified differentially included exons calculated in 1,493 RNASeq samples from the GTEx project. Each exon in each sample was categorized based on the inclusion level: low (between 0 ad 0.1), low-mid (between 0.1 and 0.5), mid-high (between 0.5 and 0.9) and high (between 0.9 and 1). Left plot represents "promoter-like" exons, right plot represents the remaining differentially included exons.

Figure S5. Transcription Factors ChIPSeq signal in "promoter-like" exons in C-higher and C-lower condition.



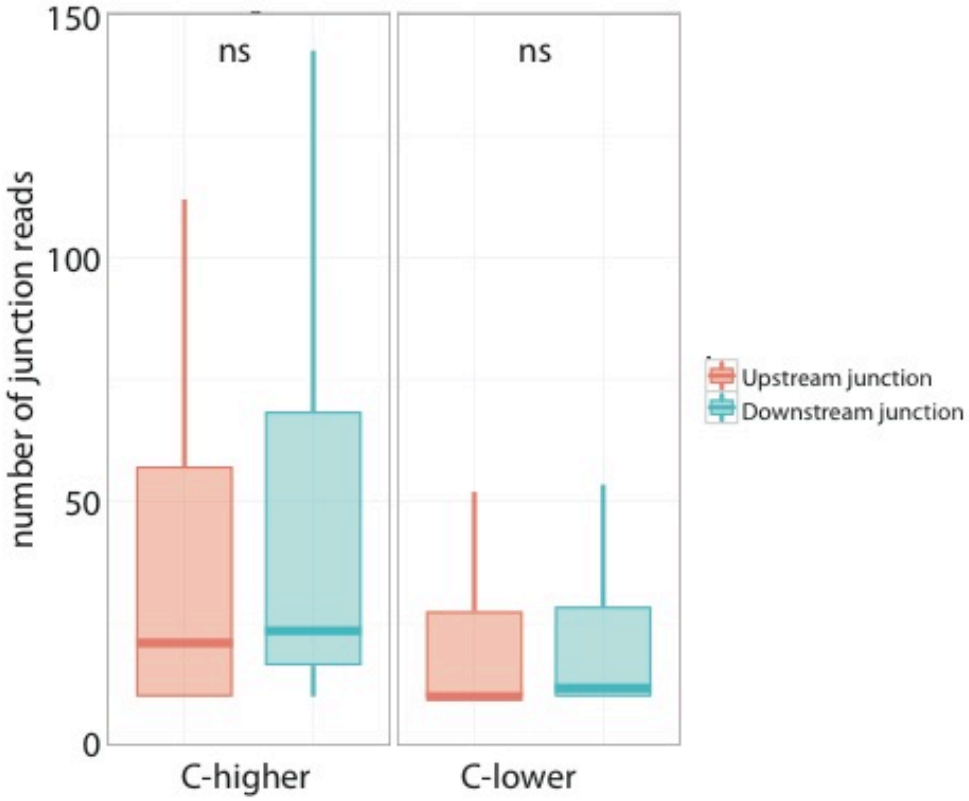
Signals are represented for "promoter-like" and flanking non-differentially included exons. Significance levels are indicated by * (0.05>p>0.01), ** (0.01>p>0.001), *** (0.001>p) and "ns" (p>0.05)

Figure S6. Dnase I sensitivity signal in regulated exons.



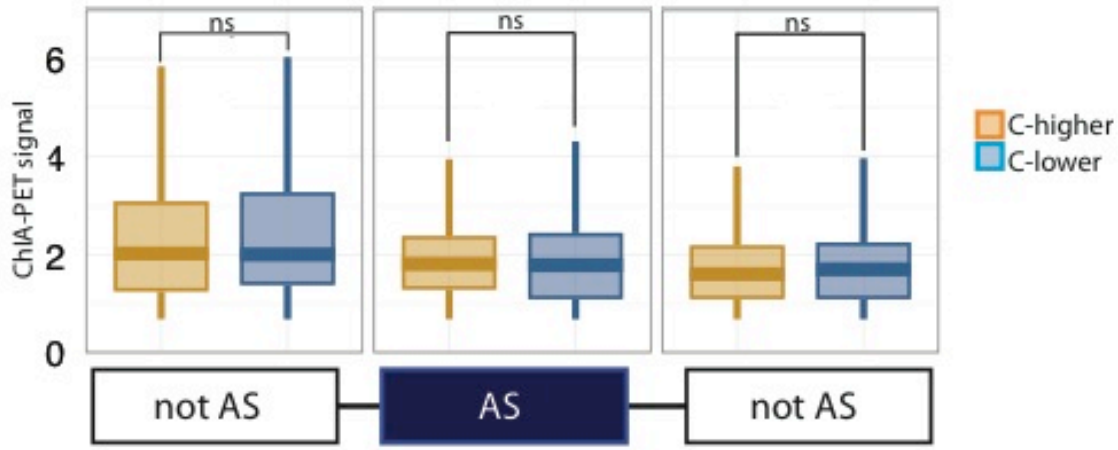
Signals are represented for regulated and flanking non-regulated exons. Significance levels are indicated by * (0.05>p>0.01), ** (0.01>p>0.001), *** (0.001>p) and "ns" (p>0.05)

Figure S7. Distribution of RNASeq junction reads in promoter like exons



Signals are represented for “promoter-like” exons in C-higher and C-lower conditions. Significance levels are indicated by * (0.05>p>0.01), ** (0.01>p>0.001) and “ns” (p>0.05)

Figure S8. ChIA-PET signal in regulated exons.



Signals are represented for regulated and flanking non-regulated exons. Significance levels are indicated by * (0.05>p>0.01), ** (0.01>p>0.001), *** (0.001>p) and "ns" (p>0.05)

Table S1. List of primers used for exon inclusion validation

Primer name	sequence
IFNGR1_dwr	GACCTGTGGCATGATCTGGT
IFNGR1_aef	GCAGATGCTTTTGAAGTCACC
IFNGR1_inclr	AACATTAGTTGGTGTAGGCACTCC
IFNGR1_skippingr	ACATTAGTTGGTGTAGGCACTGAG
RNASET2_upf	CAGCTGGCAGCGTTCTCT
RNASET2_dwr	GGCCAGTGCTGAACCATAAT
RNASET2_skipf	CAAGCGCCTGCGTGAC
Camkk2_inclf	TGCCGGAAATCAAGCTG
Camkk2_inclr	AGCGTGCAGTTCTCATCCTC
Camkk2_skipf	TGCCGGAAATCAAGATCC
Camkk2_skipr	GCTGACAGTGAGCGTTCCTC
Camkk2_dwf	TAAACGCTCCTTTGGGAACC
Camkk2_dwr	GCTGACAGTGAGCGTTCCTC
LSM7_upf	GGAGAAGAAGAAAAAGGAGAGC
LSM7_upr	TTTACCCGGATCGTCTTGTC
LSM7_inclR	GGCTAGGGCCTGTGACTCTTC
LSM7_skiplR	CAGGATTCCACTGGCTTCG
LSM7_dwf	GAATCCTGAAGGGCTTCGAC
LSM7_dwr	CTCGCATGTACTCAATGGTG
LSM7_aef	CCTCAGTGCTCACATCTCTCC
LSM7_aer	TGCAGATGTTCAGGTCCTTG
rab27a_upf	TGAAGAGGACATGTGATTGGA
rab27a_upr	ACAGGGTAGAGAACCGCTTG
rab27a_aef	TATCCACGGGCTAGCCATAC
rab27a_aer	TCCTCAGAGTGCTTCAGTGC
rab27a_dwf	TGGGAGACTCTGGTGTAGGG
rab27a_dwr	TCAATGCCCACTGTTGTGATA
rab27a_inclr	GGATAAGGGCAGAGCCTCTTTA
rab27a_skipf	CGGTTCTCTACCCTGTAAAGGTG
MUYTH_inclr	CCAGGGACCTGTATGGGTT
MUYTH_inclf	GCCAGGGACCTGTATGTAGA
MUYTH_aef	AAGTGATCTGCCCATCTTGG
MUYTH_aer	AGAACTGATAGCTCCCATGGAT
MUYTH_dwf	AAAAGGTCCCAGGTGTCCTC
MUYTH_dwr	CTGCACTGTTGAGGCTGTGT

ABI1_inclf	CAATTTTCTGCTCAGCCTCA
ABI1_inclr	GGGTGGAGCAATAGAAATTGA
ABI1_skipf	TTTTCTGCTCAGCCTCATGTT
ABI1_skipr	GGAGTTGGACTATCAGCAATTGA
ABI1_dwf	CCTCCACCAGATGACATTCC
ABI1_dwr	TGCAGCCTCCTCATCTTCAT
ERI1_upf	GCATGGAGGATCCACAGAGT
ERI1_dwr	AAGTCACTCGCACTGGAGGT
ERI1_inclf	ACGTTGTCAATCTCATCCTGAAAC
ERI1_skipR	ATTTACACTGTTGAGTTTCCTCGG
USP16_upf	ATGAGGGGATGCAGTTATGG
USP16_dwr	TGTCCGTTTCTTTCCCATGT
USP16_inclf	CTCTGTCGCCGTGGGATA
USP16_aer	GAGATCGAGGTGGGAGGAC
USP16_skipf	CTGTCGCCGTGGATTGTT
thrap3_upf	CAGCTGCGATCTCTGTGGTA
thrap3_dwr	CTGGATCCCAGACACTACCC
thrap3_inclf	TCTGTGGTAGGCCCAGTCAA
thrap3_aer	AAAACTGAGGCAGCTGGAGA
thrap3_skipf	TCTGTGGTAGGCCCAGAAGTG
SRSF6_upf	CGAGCGCGTGATCGTAGA
SRSF6_upr	GCGGCTTCCGTAGCTGTAG
SRSF6_aef	TTGTGTGACCCTTGCCCTAT
SRSF6_aer	TCTAATGGCAAAAGGCTGCT
SRSF6_dwf	AAATACGGACCACCTGTTCG
SRSF6_dwr	GCCAACTGCACCGACTAGA
SRSF6_inclr	GCCCCATTGGTCATGC
SRSF6_skipr	TCCACCTCCACCACTGC
EZH2_9Af	TGGGTATATATTGCCTGTTGGA
EZH2_9Ar	CTTCTGCAGGTGCCATTCA
EZH2_9_10f	AGTGTTACCAGCATTTGGAGG
EZH2_10r	ACGTTTTGGTGGGGTCTTTA
EZH2_9_9Af	GCGGAAGAACACAGAAACAG
EZH2_9_9Ar	TCCTAGGTAGGAGTGGCAAA

Table S2. List of cell line comparison used

Cell-pair	1	2	3	4	5
K562 vs Gm12878	YES	YES	YES	YES	YES
K562 vs HelaS3	YES	YES	YES	YES	YES
K562 vs Hepg2	YES	YES	NO	YES	YES
K562 vs Huvec	YES	YES	NO	NO	YES
Gm12878 vs HelaS3	YES	YES	YES	YES	YES
Gm12878 vs Hepg2	YES	YES	YES	YES	YES
Gm12878 vs Huvec	YES	YES	YES	NO	YES
Huvec vs HelaS3	YES	YES	YES	YES	YES
Huvec vs Hepg2	YES	YES	YES	YES	YES
HelaS3 vs Hepg2	YES	YES	YES	YES	YES

Criteria for pairwise AS exons validation	
1	AS exons should have weaker splice sites
2	AS exons should be smaller
3	Conding AS exons should be more often divisible by 3
4	Both categories should have similar gene expression
5	Both categories should have no difference in mappability

Table S3. Number of selected differential included exons

Cell-pair	More included exons	Less included exons
K562 vs Gm12878	283	227
K562 vs HelaS3	199	334
Gm12878 vs HelaS3	208	427
Gm12878 vs Hepg2	153	296
Huvec vs HelaS3	369	413
Huvec vs Hepg2	196	229
HelaS3 vs Hepg2	276	272
TOTAL	1684	2198

Table S4. List of primers used for H3K9ac validation

tag	sequence
ABI1_11f	CTCCCCCTATGCCTCAGTT
ABI1_11r	CACGAAGCCTGTGAGAGGTA
ABI1_7f	CCCCCAACAGTTCCTAATGA
ABI1_7r	GACTTCCAAGCCTAGCAGGA
EZH2_9Af	TGGGTATATATTGCCTGTTGGA
EZH2_9Ar	CTTCTGCAGGTGCCATTCA
EZH2_9f	TTTCATGCAACACCCAACAC
EZH2_9r	GGTCCACAAGGTTTGTTGTCT
camkk2_16f	ACGCTGGTCGAAGTGACTG
camkk2_16r	CAAGCTGGGAATGTGTTTGA
camkk2_12f	GCTGACTTTGGTGTGAGCAA
camkk2_12r	AGAAGATCTTGCGGGTCTCA
LSM7_aef	CCTCAGTGCTCACATCTCTCC
LSM7_aer	TGCAGATGTTCAGGTCCTTG
LSM7_upf	GGAGAAGAAGAAAAAGGAGAGC
LSM7_upr	TTTACCCGGATCGTCTTGTC

Table S5. List of RBP binding sites enriched in "promoter-like" exons

ID	RBP name	organism	db	p-value
RNCMPT00003	aret	Drosophila_melanogaster	RNAcompete	5.84E-18
RNCMPT00004	BRUNOL4	Homo_sapiens	RNAcompete	8.73E-17
RNCMPT00011	papi	Drosophila_melanogaster	RNAcompete	7.60E-15
RNCMPT00027	HNRNPL	Homo_sapiens	RNAcompete	2.15E-10
RNCMPT00051	RBM38	Homo_sapiens	RNAcompete	2.15E-10
RNCMPT00069	sm	Drosophila_melanogaster	RNAcompete	2.15E-10
RNCMPT00113	RBM4	Homo_sapiens	RNAcompete	2.15E-10
RNCMPT00114	aret	Drosophila_melanogaster	RNAcompete	2.15E-10
RNCMPT00134	SRSF4	Homo_sapiens	RNAcompete	2.15E-10
RNCMPT00150	ESRP2	Homo_sapiens	RNAcompete	2.15E-10
RNCMPT00160	HNRNPH2	Homo_sapiens	RNAcompete	2.15E-10
RNCMPT00166	CELF3	Homo_sapiens	RNAcompete	2.15E-10
RNCMPT00178	hnRNPLL	Homo_sapiens	RNAcompete	6.66E-04
RNCMPT00179	Rbm24	Mus_musculus	RNAcompete	8.08E-04
RNCMPT00216	RRM_1	Trypanosoma_brucei	RNAcompete	1.01E-03
RNCMPT00246	Pcbp4	Mus_musculus	RNAcompete	3.33E-03
RNCMPT00270	aret_CONSTRUCT	RNAcompete_CONSTRUCTS	RNAcompete	2.89E-02
RNCMPT00283	Rbm24	Mus_musculus	RNAcompete	4.20E-02
RNCMPT00285	Rbm24	Mus_musculus	RNAcompete	4.20E-02

Table S6. List of Transcription factor binding sites enriched in "promoter-like" exons

ID	TF name	organism	db	p-vavlue
MA0472.1	EGR2	Mus_musculus	JASPAR core	0.00E+00
MA0592.1	ESRRA	Homo_sapiens	JASPAR core	3.69E-04
MA0599.1	KLF5	Homo_sapiens	JASPAR core	6.15E-03
MA0491.1	JUND	Homo_sapiens	JASPAR core	1.23E-02
MA0477.1	FOSL1	Homo_sapiens	JASPAR core	2.82E-02
MA0131.1	HINFP	Homo_sapiens	JASPAR core	3.74E-02
MA0470.1	E2F4	Homo_sapiens	JASPAR core	3.74E-02